Statistics
in Medicine

Tutorial in Biostatistics

Received 22 July 2010, Accepted 30 October 2012 Published online 3 December 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5686

Methods for dealing with time-dependent confounding

R. M. Daniel,[a*†] S. N. Cousens,[a] B. L. De Stavola,[a] M. G. Kenward[a] and J. A. C. Sterne[b]

Longitudinal studies, where data are repeatedly collected on subjects over a period, are common in medical research. When estimating the effect of a time-varying treatment or exposure on an outcome of interest measured at a later time, standard methods fail to give consistent estimators in the presence of time-varying confounders if those confounders are themselves affected by the treatment. Robins and colleagues have proposed several alternative methods that, provided certain assumptions hold, avoid the problems associated with standard approaches. They include the g-computation formula, inverse probability weighted estimation of marginal structural models and g-estimation of structural nested models. In this tutorial, we give a description of each of these methods, exploring the links and differences between them and the reasons for choosing one over the others in different settings. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: time-dependent confounding; g-computation formula; inverse probability weighting; g-estimation; marginal structural model; structural nested model

1. Introduction

1.1. Motivation

In many longitudinal medical studies, patients' treatment or exposure (henceforth referred to as treatment) changes over time and is measured several times during the study, along with other time-changing covariates. For example, type II diabetes patients recruited into a study comparing two antiglycaemic drugs may be followed up on several occasions, on each of which their HbA_{1c} (a long-term measure of blood glucose level), blood pressure, cholesterol level, body mass index, anaemia status, and others variables are measured, and changes may be made to the dose and type of drug prescribed to them. Suppose that, in such a setting, we wish to compare the effect of the two treatments on HbA_{1c} 18 months after recruitment and on the risk (or hazard) of experiencing a cardiac event in the 18 months following recruitment. Even if initial treatment is determined at random, it is likely that the study protocol will allow, for ethical reasons, for the dose and type of treatment to be changed according to the current (and past) values of HbA_{1c} and other covariates. A high HbA_{1c} indicating poor control would likely lead to increasing the dose of the current drug, or switching to a different drug. But high HbA_{1c} is also thought to lead to an increased risk of a cardiac event, making HbA_{1c} at a particular time a confounder of the relationship between subsequent treatment and the outcome. Because HbA_{1c} varies over time (in a way that cannot be foreseen at baseline), it is called a *time-varying confounder*. To estimate the causal effect of treatment on risk of cardiac event, it seems necessary to control for HbA_{1c} in the analysis. However, not only does HbA_{1c} affect treatment but also the reverse is true. An effective antiglycaemic drug lowers HbA_{1c}, and thus the current value of the treatment variable has a causal effect on future values of HbA_{1c}. This means that controlling for HbA_{1c} is problematic, because future measurements of HbA_{1c} lie on the causal pathway between past treatment and the outcome, and thus conditioning on

[a] *Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, London, U.K.*
[b] *Department of Social Medicine, University of Bristol, Bristol, U.K.*
**Correspondence to: Rhian Daniel, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, U.K.*
[†] *E-mail: rhian.daniel@lshtm.ac.uk*

HbA_{1c} blocks some of the effect of the treatment and, in addition, conditioning on a consequence of treatment risks inducing collider-stratification bias (Section 3).

1.2. Tutorial aims and outline

Robins [1] was the first to suggest a method, the g-computation formula, for estimating the causal effect of a time-varying treatment in the presence of time-varying confounders that are affected by treatment. He and colleagues have proposed two further methods to be used in this setting: inverse probability weighted estimation of marginal structural models (MSMs) [2] and g-estimation of structural nested models (SNMs) [3].

Robins and Hernán [4] have written an excellent tutorial on this topic. Our aim in this paper is to provide a more accessible introduction to these three methods and the links and differences between them, by demonstrating their use in simple worked examples and giving computer code for their implementation. Many important issues (such as the comparison of dynamic treatment regimes, see Section 2.7) are not covered here, but are covered by Robins and Hernán [4]. We therefore highly recommend the chapter by Robins and Hernán as a sequel to this article.

We start, in Section 2, with the notation and key definitions and concepts used in the remainder of the article. This is followed, in Section 3, by an introduction to the problem of time-dependent confounding. In Section 4, we introduce both the g-computation formula and inverse probability weighted estimation of MSMs in a nonparametric context, and in Section 5, we introduce their parametric extensions. We introduce g-estimation of SNMs in Section 6. In Section 7, we provide some extensions of the approaches described here, followed, in Section 8, by a comparison of the three approaches, discussing the reasons for choosing one method over the others in various settings. Section 9 summarises the paper. We provide the STATA code for all the examples, along with further examples and additional figures, in the Supporting Information (available from the journal web page).[‡]

2. Setting, notation and definitions

2.1. The setting

We consider a setting in which n subjects, labelled $i = 1, \ldots, n$, enter a study at baseline (time τ_0) and are subjected to treatment $A_{0,i}$. In a clinical trial, the value of $A_{0,i}$ is determined by the study protocol, usually at random. In an observational study, $A_{0,i}$ is merely observed and recorded. Often, $A_{0,i}$ is binary (treatment or control), but more generally, we say that $A_{0,i}$ takes a value in the set $\mathcal{A}_0$. A collection of covariates, $L_{0,i}$, is also measured at baseline, taking values in the set $\mathcal{L}_0$. In keeping with the literature in this area, we do not use a bold typeface for $L_{0,i}$, even though it is in general a vector.

There are $T+1$ subsequent follow-up visits, labelled $t = 1, \ldots, T+1$, occurring at times $\tau_1, \ldots, \tau_{T+1}$, respectively. At visit t, $A_{t,i}$ (taking a value in the set $\mathcal{A}_t$) and $L_{t,i}$ (taking values in the set $\mathcal{L}_t$) are recorded. We assume that $A_{t,i}$ is the value taken by the treatment in the time interval $[\tau_t, \tau_{t+1})$ and similarly that the covariate values $L_{t,i}$, measured just before $A_{t,i}$, remain unchanged during this interval.

Note that $L_{0,i}$ could contain time-fixed covariates (e.g. age[§] and gender) as well as baseline measurements of the time-varying covariates.

An outcome Y_i is observed for each subject, measured at the final visit $T+1$. Y_i can be continuous or binary. The setting in which Y_i is the time to some event of interest is not considered here but is discussed elsewhere in the literature [3, 5–8].

We assume the n study subjects to have been selected independently and at random from a much larger population of N individuals that can be taken to be infinite; in some settings, such a population may be hypothetical. Our statistical and epidemiological task is to make inference about the causal effect (to be defined in Section 2.4) of the time-varying treatment on the outcome in this population, using the observational data on the n sampled subjects.

[‡]*Supporting information may be found in the online version of this article.*

[§]*Age is a time-fixed covariate because, even though it varies over time, it does so in a way that is entirely deterministic. The age of a subject at visit* t *is known at baseline.*

2.2. Treatment and covariate histories

We use $\bar{A}_{t,i} = (A_{0,i}, \ldots, A_{t,i})$ to denote the treatment history and $\bar{L}_{t,i} = (L_{0,i}, \ldots, L_{t,i})$ the covariate history of subject i up to and including visit t. $\bar{A}_{t,i}$ therefore takes a value in the set $\bar{\mathcal{A}}_t = \mathcal{A}_0 \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_t$, and $\bar{L}_{t,i}$ takes a value in the set $\bar{\mathcal{L}}_t = \mathcal{L}_0 \times \mathcal{L}_1 \times \cdots \times \mathcal{L}_t$. We use $\bar{A}_i$ and $\bar{L}_i$ for the entire history of A and L up to and including visit T, respectively, for subject i. That is, we write $\bar{A}_i = \bar{A}_{T,i}$ and $\bar{L}_i = \bar{L}_{T,i}$. Similarly, we write $\bar{\mathcal{A}}$ for $\bar{\mathcal{A}}_T$ and $\bar{\mathcal{L}}$ for $\bar{\mathcal{L}}_T$.

We use capital letters to denote random variables and lower case letters to denote their realised values. Thus, we use $\bar{a} \in \bar{\mathcal{A}}$ to denote a generic realised treatment history and $\bar{l} \in \bar{\mathcal{L}}$ to denote a generic realised covariate history.

2.3. Potential outcomes and counterfactuals

Along with the *observed outcome* Y_i for each subject, for each possible realisation of the treatment history $\bar{a} \in \bar{\mathcal{A}}$, we also define $Y_i^{\bar{a}}$, the *potential outcome* that would have been observed had the subject, possibly contrary to fact, received treatment history $\bar{a}$.

For $Y_i^{\bar{a}}$ to be well defined, we require that setting the treatment of another subject j to $\bar{b}$ would have no effect on the potential outcome for subject i. This is implicit in the fact that only subject i's hypothetical treatment trajectory $\bar{a}$ is included in the superscript for $Y_i^{\bar{a}}$. This assumption is known as *no interference* [9], and we will make this assumption throughout.[¶] Making any sensible inference about $Y^{\bar{a}}$ from the observational data on Y and A (and covariates) will also require the *consistency assumption*, namely that the actual outcome is equal to the potential outcome under assignment to the treatment actually taken, that is, that $Y_i = Y_i^{\bar{A}_i}$ [10].

Under the consistency assumption, exactly one potential outcome is observed: $Y_i^{\bar{A}_i}$. All other potential outcomes are called *counterfactual*. Rubin [9] formalised the notion of potential outcomes, and Robins [1] extended it to accommodate time-varying treatments.

2.4. Causal effects

Potential outcomes allow us to define what it means for $\bar{A}$ to have a *causal effect* on Y and furthermore to characterise this causal effect in terms of the parameters of a structural model.

2.4.1. The existence of a causal effect. If $Y_i^{\bar{a}}$ is the same for all $\bar{a} \in \bar{\mathcal{A}}$ and this holds for each subject i in the population,[‖] then we say that there is *no causal effect* of $\bar{A}$ on Y. If there exists at least one subject i in the population and at least two distinct treatment histories $\bar{a}$ and $\bar{b}$ in $\bar{\mathcal{A}}$ such that $Y_i^{\bar{a}} \neq Y_i^{\bar{b}}$, then there is a *causal effect* of $\bar{A}$ on Y.

2.4.2. Characterising the causal effect. In its fullest form, the *causal effect* of $\bar{A}$ on Y is given by the distribution of $Y^{\bar{a}}$ in the population for every $\bar{a} \in \bar{\mathcal{A}}$.[**] Often, we will focus on one aspect of this distribution, and in this tutorial, this will typically be the mean, $E\left(Y^{\bar{a}}\right)$. Note however that for right-censored time-to-event outcomes, the aspect of interest would often instead be the survivor function, $S_{Y^{\bar{a}}}(\tau) = P\left(Y^{\bar{a}} > \tau\right)$.

2.4.3. Joint, total and direct causal effects. Each potential outcome $Y^{\bar{a}}$ can be conceptualised as having arisen from a *hypothetical intervention*. It is the value that Y would have taken had we intervened on $\bar{A}$ and set it to $\bar{a}$. Throughout this tutorial, we consider hypothetical *joint* interventions on all of $\bar{A}$.

Contrast this with a situation in which we only intervene on, say, A_0, setting its value to a_0, but allow all future treatments to attain their natural values given this earlier intervention. The distribution of Y^{a_0} for different values of $a_0 \in \mathcal{A}_0$ would then tell us about the *total* effect of A_0 on Y. The term 'total' indicates that it includes the effect of later treatments on Y in so much as these have been affected by A_0.

[¶] *It is clearly violated in certain settings, for example, if the treatment (A) is vaccination against acquiring an infectious disease (Y), where the hypothetical immunity status of others affects a particular subject's potential outcome.*

[‖] *Note that, in a slight abuse of notation, i is used here to index the subjects in the population rather than the sample.*

[**] *More precisely, writing $\mathcal{Y}$ for the support of Y, the causal effect of $\bar{A}$ on Y is a mapping from $\bar{\mathcal{A}} \times \mathcal{Y}$ to $\mathbb{R}^+$ for continuous Y, and from $\bar{\mathcal{A}} \times \mathcal{Y}$ to $[0, 1]$ for discrete Y, where the mapping, for each value $\bar{a}$ of $\bar{\mathcal{A}}$ and each y in $\mathcal{Y}$, gives the value of the probability density/mass function of $Y^{\bar{a}}$ evaluated at y.*

Suppose that the treatment has a large but short-term effect on the outcome, and also that treatment is rarely ceased once it is started. Then, the *total* effect of A_0 would be large, because a change in A_0 leads to a change in $A_1, \ldots,$ which leads to a change in A_T, which has a large effect on Y. In this setting the distribution of Y^{a_0} would differ for different values of a_0. However, the distribution of $Y^{\bar{a}}$ could still be the same for all values of a_0, indicating that the effect of A_0 on Y is entirely *mediated* by later treatments, with no *direct effect* of A_0.

The *joint* effect of $\bar{A}$ on Y consists of a collection of *direct* effects: (1) the direct effect of A_0 on Y unmediated by $\{A_1, \ldots, A_T\}$; (2) the direct effect of A_1 on Y unmediated by $\{A_2, \ldots, A_T\}$; and so on. The fact that the joint effect is a collection of direct effects (as opposed to a collection of total effects) will be crucial to our appreciation of the limitation of standard methods in this context, as discussed in Section 3.3. Also, it highlights the strong link between the setting discussed in this tutorial and the estimation of direct and indirect effects in the exploration of causal mechanism—for a review of the large body of recent literature on 'causal mediation analysis', see [11].

2.4.4. More parsimonious characterisations: structural models. Suppose, as suggested in Section 2.4.2, that we are interested in looking at the causal effect of $\bar{A}$ on the mean of Y. Even if treatment were binary, there are 2^{T+1} values of $\bar{a}$, and hence the causal effect is characterised by 2^{T+1} quantities:

$$\left\{E\left(Y^{\bar{a}}\right) : \bar{a} \in \bar{\mathcal{A}}\right\}. \tag{1}$$

As T increases, the high-dimensional nature of this characterisation leads to difficulties both with estimation (due to an insufficient number of subjects following any given trajectory) and with interpretation (due to too many potential comparisons).

Often, therefore, we may opt for a more parsimonious characterisation of the causal effect via a structural model. For example, we may posit that

$$E\left(Y^{\bar{a}}\right) = \varphi^{-1}\left(\varrho + \varsigma \sum_{t=0}^{T} a_t\right), \tag{2}$$

where $\varphi(\cdot)$ is a link function, such as log, logit or the identity. By making the (possibly incorrect) assumption that the effect of treatment is exactly cumulative on the scale of the link function, the 2^{T+1} quantities in (1) have been reduced to two parameters, ϱ and ς. The model implied by (2) is an example of an *MSM*. We provide later in the article more details of such models and their parameters.

2.4.5. Use of the term 'causal effect'. In this tutorial, we will use *causal effect* rather loosely to mean a parameter or parameters arising from a structural model, as well as the full characterisation given in Section 2.4.2. The exact meaning in any given situation will be evident from the context.

2.5. Identifiability

Whether our target of estimation is $E\left(Y^{\bar{a}}\right)$ (or some other aspect of the distribution of $Y^{\bar{a}}$) or a derived summary smoothed across many different values of $\bar{a}$ in the form of a parameter from a structural model, the inferential task is made difficult by the simple fact that $Y^{\bar{a}}$ is observed on at most a subset of our n study subjects, namely those with $\bar{A} = \bar{a}$. In this sense, all causal inference problems are missing data problems, and assumptions are needed in order to proceed.

Consider, for example, a very large double-blind randomised controlled trial, with no missing data, measurement error or non-adherence, and where each subject is randomised at baseline to receive one treatment trajectory out of all 2^{T+1} possible trajectories in $\bar{\mathcal{A}}$.

Even in such an idealised setting, it of course remains the case that $Y^{\bar{a}}$ is only observed for a subset of the study subjects. However, it would be reasonable in this situation to take the average of $Y^{\bar{a}}$ for those with $\bar{A} = \bar{a}$ as an estimator of $E\left(Y^{\bar{a}}\right)$ for the whole population, because randomisation should guarantee that the subset with $\bar{A} = \bar{a}$ is representative of (or *exchangeable* with) the study sample and hence of the population.

In this case, we would say that the causal effect is *identified* under the exchangeability assumption made plausible by randomisation, by which we mean that it can be consistently estimated from the observational data under this assumption.

Identification is possible under a weaker assumption, which we state in the next section, but—as we would expect—identifiability under this weaker assumption requires more complicated estimation methods, and these are the topic of this tutorial.

2.6. The 'no unmeasured confounding' assumption

In addition to 'no interference' and consistency, a sufficient assumption under which causal effects are identifiable in the setting described previously is the *sequentially ignorable treatment assignment* assumption, also known as the *no unmeasured confounding* or *conditional exchangeability* assumption. Formally, it is the assumption that[††]

$$Y^{\bar{a}} \perp\!\!\!\perp A_t \,|\, \bar{L}_t, \bar{A}_{t-1} \quad \forall \bar{a} \in \bar{\mathcal{A}}, \forall t \in \{0, \ldots, T\}.$$

That is, conditional on treatment history up to visit $t-1$ and the history of all measured covariates up to visit t, the treatment received at visit t is independent of the potential outcomes. Under this assumption, conditional on $\bar{L}_t$ and $\bar{A}_{t-1}$, treated and untreated subjects at visit t are exchangeable in the sense that were they all to remain untreated, say, the distribution of Y would not differ between the two groups. Rubin [9] first stated this important assumption for the single timepoint setting, and Robins [1] extended this to time-varying treatments.

Note that standard analyses (i.e. in the present context, those that adjust for $\bar{L}$ in a regression model), although often less explicit about the assumptions being made, implicitly require a 'no unmeasured confounding' for a causal interpretation to be justified. In fact, standard analyses assume *more* than this, because they additionally require that the time-varying confounders are not affected by the treatment. Similarly, although not always made explicit, causal inference from standard analyses also require no interference and consistency.

2.7. Static and dynamic regimes

The hypothetical interventions discussed thus far (essentially setting $\bar{A}$ to a set of constant values $\bar{a}$) are known as *static* regimes, because, in the hypothetical interventions being considered, treatment does not depend on the values of the time-varying covariate. In many situations, the practical question of interest requires a comparison of *dynamic* regimes, such as 'treat until patient becomes anaemic, then stop treatment'. In the latter case, the treatment regime is a function of the covariate history. In the study of chronic diseases such as HIV and diabetes, where effective treatments are well established but their optimal administration may not be known, relevant clinical questions are indeed more likely to be based on a comparison of dynamic regimes than on a comparison of static regimes. We can use all three methods discussed in this tutorial to compare dynamic regimes, and we provide relevant references in Section 7. In the interest of simplicity, we focus only on the comparison of static regimes.

2.8. Causal diagrams

A *causal diagram* $\mathcal{G}$ consists of nodes (or points) denoting variables, and arrows between nodes denoting the assumed direction of causal influence. Any variable that is the common cause of two or more variables in $\mathcal{G}$ must itself be in $\mathcal{G}$. Figure 1 shows examples of causal diagrams. Arrows are permitted to be null: an arrow can exist when in fact a causal influence is absent; it is thus the absence of arrows that encode causal assumptions.

If there is an arrow from A to B in $\mathcal{G}$, then A is said to be a *parent* of B, and B a *child* of A. A *path* from one node to another is a set of one or more arrows (in any direction) connecting the two nodes. In Figure 1(a), for example, there are two paths from A_0 to Y: $A_0 \rightarrow Y$ and $A_0 \leftarrow L_0 \rightarrow Y$. Another important concept is that of a *collider*, that is a node on a path at which two arrowheads collide. For example, L_1 is a collider on the path $A_0 \rightarrow L_1 \leftarrow U_0 \rightarrow Y$ in Figure 1(e). A path from A to B in which all arrows point 'forwards' (away from A and towards B) is called a *directed* path. If there is a directed path from A to B, then A is an *ancestor* of B, and B is a *descendant* of A.

Colliders (common effects) are important because they do not transmit (marginal) associations;[‡‡] two variables are (marginally) independent if they are not linked by a path or if every path between them

[††] *By convention,* $\boldsymbol{A}_{-1} = \mathbf{0}$.

[‡‡] *To understand why, note that two variables can only be associated if either one variable causes the other, or if both share a common cause; this cannot be the case when every path between them contains a collider.*

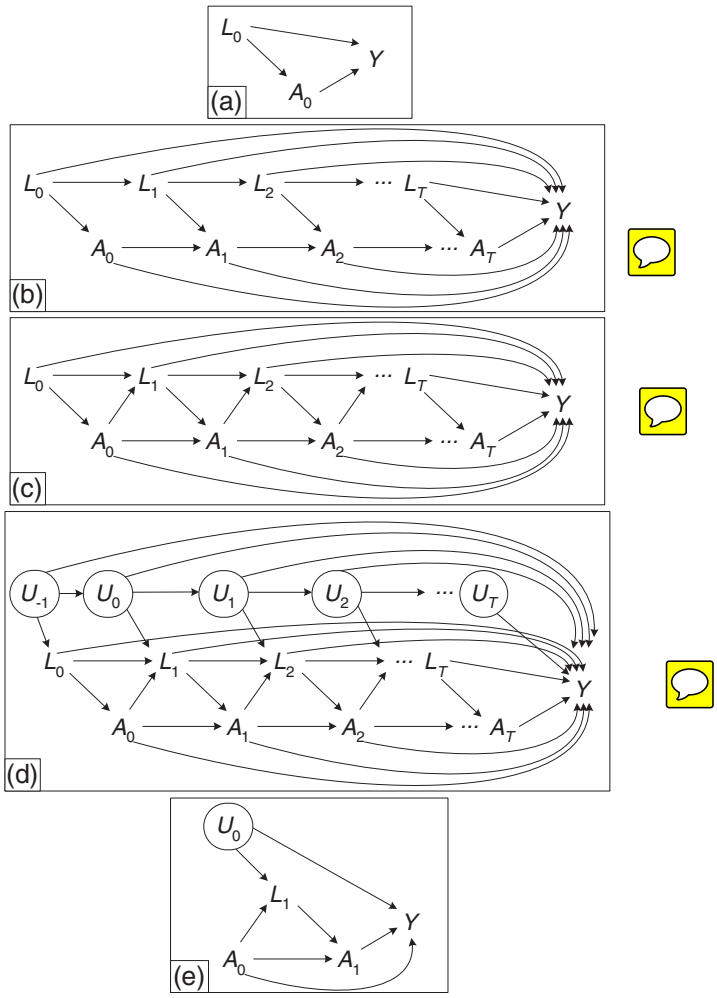

Figure 1. Causal diagrams showing (a) time-fixed common cause confounding, (b) time-dependent confounding, (c) time-dependent confounding with confounder affected by past treatment, (d) the same as (c) but with the confounder–outcome relationship not necessarily purely causal and (e) our simplified example.

contains at least one collider. However, conditioning on a collider (or any of its descendants) induces a conditional association between the parents of the collider in the graph, even if these parents are marginally independent.[§§]

A path p from A to B is said to be *closed with respect to a set of variables* $\mathcal{S}$ if (1) p contains a collider C, where neither C nor any of its descendants are in $\mathcal{S}$, and/or (2) p contains a non-collider D, which is in $\mathcal{S}$. If every path from A to B is closed by $\mathcal{S}$, then A and B are conditionally independent given $\mathcal{S}$. Thus, informally, we see that there are two ways of 'blocking' associations: by conditioning on non-colliders and by not conditioning on colliders (or their descendants).

Any path that is not closed with respect to $\mathcal{S}$ is said to be open (with respect to $\mathcal{S}$).

A path from A to B that starts with an arrow into A is known as a *back-door* path. The existence of an open (with respect to $\emptyset$) back-door path from an exposure of interest to an outcome (such as $A_0 \leftarrow L_0 \rightarrow Y$ in Figure 1(a)) shows that a naïve analysis looking for a marginal association between exposure and outcome cannot be given a causal interpretation; in the setting of Figure 1(a), the causal path $A_0 \rightarrow Y$ is confounded by the back-door path $A_0 \leftarrow L_0 \rightarrow Y$.

[§§] *For some intuition on why this is the case, consider a selective school in which pupils are accepted on the basis of academic and sporting ability. Suppose, for the sake of this example, that academic and sporting ability are independent in the population. Within the selective school, however, academic and sporting ability are (negatively) associated: a pupil selected at random found to have poor sporting ability must be academically able.*

Associations are transmitted along open paths but can be blocked by conditioning on a variable along that path (e.g. by conditioning on L_0 in Figure 1(a)). Associations are not transmitted along paths containing a collider unless we condition on it (or one of its descendants). These two observations form the backbone of causal diagrams and their use in causal inference.

For more details, see [12, 13].

3. Time-dependent confounding: an introduction to the problem

3.1. Time-fixed confounding

Suppose that $T = 0$, and thus the only variables are A_0, L_0 and Y. If L_0 lies on a back-door path from A_0 to Y, then L_0 is a confounder of the relationship between A_0 and Y [14]. Even if there were no causal effect of A_0 on Y, a crude comparison of Y for different values of A_0 would show an association between A_0 and Y whenever both the $L_0 \rightarrow A_0$ and $L_0 \rightarrow Y$ arrows are non-null. In the presence of a causal effect of A_0 on Y, the crude estimator is, in general, biased (and inconsistent) for this causal effect.

For example, if patients with a high $\mathrm{HbA_{1c}}$ at baseline are more likely to be selected into the treatment group and if a high $\mathrm{HbA_{1c}}$ leads to an increased risk of a cardiac event, then if there is no causal effect of treatment on outcome, a naïve analysis would suggest that treatment is harmful, because those in the treatment group tend to have a higher risk of experiencing a cardiac event. However, if we repeated the analysis controlling for L_0, for example, in a logistic regression with cardiac event Y as the binary outcome and treatment A_0 and $\mathrm{HbA_{1c}}$ as explanatory variables, then we would see no (conditional) association between A_0 and Y.[¶¶] More generally, the conditional causal effect of A_0 on Y, given L_0, could be consistently estimated from such a regression model, if correctly specified.

This is an example of time-fixed confounding, and the causal diagram in Figure 1(a) illustrates this.

3.2. Time-dependent confounding

Suppose that $T > 0$, and at each visit t, $A_{t,i}$ is causally influenced by the previous value of treatment, $A_{t-1,i}$, and the current value of the covariates, $L_{t,i}$. Then, suppose that Y_i is causally influenced by the whole histories $\bar{A}_i$ and $\bar{L}_i$. In a crude analysis that ignores $\bar{L}$, the effect of $\bar{A}$ on Y is confounded: L is a time-varying confounder. The causal diagram in Figure 1(b) illustrates this situation. We could also have included arrows from $\bar{A}_{t-2}$ and $\bar{L}_{t-1}$ to A_t and from $\bar{L}_{t-2}$ to L_t, and so on; these have been omitted to make the diagram more readable. Note, however, that there are no arrows from $\bar{A}_t$ to L_{t+1}; these omissions are crucial. Controlling for L in an appropriately specified regression model with Y as the outcome and $\bar{A}$ and $\bar{L}$ as explanatory variables, would, in this situation, give a consistent estimator of the conditional causal effect of $\bar{A}$ on Y given $\bar{L}$, as in the time-fixed setting.

3.3. Time-dependent confounding where the confounder is affected by the treatment

In Figure 1(c), arrows have been added from A_{t-1} to L_t; that is, the time-varying covariate is now affected by past treatment. For example, it is realistic to assume that clinicians are more likely to prescribe an antiglycaemic drug when $\mathrm{HbA_{1c}}$ is high and that high $\mathrm{HbA_{1c}}$ can lead to an increased risk of a cardiac event. It is also known that antiglycaemic drugs work by lowering $\mathrm{HbA_{1c}}$, and thus the causal diagram shown in Figure 1(c) seems reasonable. L_t remains a confounder of the relationship between A_t and Y, but L_t is now also on the causal pathway between $\bar{A}_{t-1}$ and Y. Controlling for $\bar{L}$ in the way described in Section 3.2 is no longer satisfactory, because this blocks (i.e. does not include) any of the effect of $\bar{A}_{t-1}$ acting through $\bar{L}$. In the case of this example, it is believed that most of the effect of treatment is through lowering $\mathrm{HbA_{1c}}$, and so controlling for $\bar{L}$ could introduce serious bias.

It is also possible that unmeasured factors U_{t-1} in the time interval $[\tau_{t-1}, \tau_t)$ influence both $\mathrm{HbA_{1c}}$ at visits t and Y. For example, if information on a subject's diet is not collected, then it would likely be an unmeasured time-varying common cause of both $\bar{L}$ and Y. Figure 1(d) has been augmented to reflect this, with the circles surrounding $\bar{U}$ indicating that they are unmeasured. This introduces another problem associated with controlling for $\bar{L}$. Because U_{t-1} and A_{t-1} are both causes of L_t (i.e. L_t is a collider on the path $A_{t-1} \rightarrow L_t \leftarrow U_{t-1} \rightarrow Y$ as discussed in Section 2.8), conditioning on L_t induces

[¶¶] *This, of course, assumes that the logistic regression model for Y given HbA_{1c} has been correctly specified. If $\mathrm{logit}\{E(Y)\}$ were a function of $\log(HbA_{1c})$, say, then some residual confounding would remain.*

an association between U_{t-1} and A_{t-1} and hence between A_{t-1} and Y even if A_{t-1} has no causal effect on Y. This is an example of what Hernán *et al.* [14] call *collider-stratification bias*. Particularly in settings with high-dimensional L_t, the existence of such a U_{t-1} is very likely.

Note that the *total* causal effect of A_t on Y could be estimated, for each t, even in the scenario shown in Figure 1(c), using standard adjustment for $(\bar{A}_{t-1}, \bar{L}_t)$. Such effects would, however, include the effect of A_t mediated by $(A_{t+1}, \ldots, A_T)$. We refer the reader back to the discussion in Section 2.4.3. For consideration of the *joint* causal effect of A on Y—the topic of this tutorial—standard methods are problematic.

3.4. A simple simulated example illustrating the problem with standard methods

We consider the simplest possible scenario in which the problem of time-dependent confounding with the confounder affected by past treatment can be illustrated. The setting we consider is similar to the one discussed by Robins and Wasserman [15] and Robins and Hernán [4]. It is a simplified version of the causal diagram in Figure 1(d) with $T = 1$, with treatment initially determined at random (no L_0), univariate L_1 and U_0, and no causal effect of L_1 on Y (except via A_1). Figure 1(e) shows the simplified diagram.

For example, let A_0 and A_1 be binary indicators for the prescription of antiglycaemic drug at visits 0 and 1, respectively. So that the arrow from L_1 to Y can be justifiably removed (purely for simplicity), we take L_1 to be an indicator for anaemia at visit 1 (measured just before A_1 is chosen) and Y becomes log HbA_{1c} measured at visit 2. U_0 is an unmeasured binary indicator, with $U_0 = 1$ if the subject is healthy prior to randomisation and $U_0 = 0$ if the subject is unhealthy. Some antiglycaemic drugs are thought to cause anaemia [16], and thus clinicians are less likely to prescribe antiglycaemic drugs to anaemic patients. Poor health in general is likely to lead to an increased risk of anaemia as well as an increased HbA_{1c}. The causal diagram of Figure 1(e) encapsulates this causal structure. We are interested in estimating the (joint) causal effect of (A_0, A_1) on Y.

3.4.1. Simulated dataset I. We simulate data on a sample of 2000 subjects as follows:

- U_0 is drawn from a Bernoulli distribution with mean 0.4.
- A_0 is drawn from a Bernoulli distribution with mean 0.5, that is, initial treatment is randomised.
- Conditional on U_0 and A_0, L_1 is drawn from a Bernoulli distribution with $P(L_1 = 1) = 0.25 + 0.3A_0 - 0.2U_0 - 0.05A_0U_0$. That is, anaemia at visit 1 is more common among those treated at visit 0 and among those who were unhealthy at visit 0. There is a small interaction effect, indicating that the side effect of treatment at visit 0 is more pronounced in unhealthy individuals.
- Conditional on A_0 and L_1, A_1 is drawn from a Bernoulli distribution with $P(A_1 = 1) = 0.4 + 0.5A_0 - 0.3L_1 - 0.4A_0L_1$, that is, being treated at visit 0 increases the probability of being treated at visit 1, whereas being anaemic at visit 1 decreases the probability of being treated at visit 1, with this effect more pronounced among those who were treated at visit 0.
- Finally, conditional on U_0, A_0 and A_1, log HbA_{1c} (Y) is generated from a normal distribution with mean $2.5 - 0.5A_0 - 0.75A_1 + 0.2A_0A_1 - U_0$ and standard deviation 0.2. This means that there is a beneficial causal effect of treatment at either visit, with the two effects combining subadditively. Conditional on treatment, HbA_{1c} is also higher on average for the unhealthy patients.
- Because the aforementioned description is of the data generating process, the parameters are all structural (i.e. causal). In other words, the counterfactual means are given by averaging over the distribution of U_0:

$$E\left\{Y^{(a_0,a_1)}\right\} = 2.5 - 0.4 - 0.5a_0 - 0.75a_1 + 0.2a_0a_1 = 2.1 - 0.5a_0 - 0.75a_1 + 0.2a_0a_1. \quad (3)$$

  Equivalently, given $U_{0,i}$, the potential outcomes $Y_i^{(a_0,a_1)}$ can be thought of as being generated independently for each (a_0, a_1) and for each i from a normal distribution with mean $2.5 - 0.5a_0 - 0.75a_1 + 0.2a_0a_1 - U_{0,i}$ and standard deviation 0.2.

Note that the data generating models for A_0 and A_1 do not include U_0. Thus, $A_0 \perp\!\!\!\perp U_0$ and $A_1 \perp\!\!\!\perp U_0 \,|L_1$. Because, conditional on U_0, the variation in $Y^{(a_0,a_1)}$ is independent of all other random variables, this means that $A_0 \perp\!\!\!\perp \left\{Y^{(a_0,a_1)}\right\}$ and $A_1 \perp\!\!\!\perp \left\{Y^{(a_0,a_1)}\right\} | A_0, L_1$, that is, the data have been generated under the 'no unmeasured confounding' assumption.

We provide the STATA code used to simulate this dataset in Section A1 of the Supporting Information, and the dataset is available from the corresponding author upon request. Of the 2000 subjects in the simulated dataset, 664 are assigned to the treatment trajectory $(A_0, A_1) = (0, 0)$, 440 to $(1, 0)$, 352 to $(0, 1)$ and 544 to $(1, 1)$. Of the 2000 subjects, 640 are diagnosed with anaemia at visit 1 ($L_1 = 1$).

The expected values of the four potential outcomes, $Y^{(0,0)}$, $Y^{(1,0)}$, $Y^{(0,1)}$ and $Y^{(1,1)}$, are obtained from (3) giving $E\left\{Y^{(0,0)}\right\} = 2.1$, $E\left\{Y^{(1,0)}\right\} = 1.6$, $E\left\{Y^{(0,1)}\right\} = 1.35$ and $E\left\{Y^{(1,1)}\right\} = 1.05$.

We may also be interested in the parameters of an equation such as

$$E\left\{Y^{(a_0,a_1)}\right\} = \gamma_{\text{int}} + \gamma_0 a_0 + \gamma_1 a_1 + \gamma_{01} a_0 a_1, \tag{4}$$

which are related to $E\left\{Y^{(0,0)}\right\}$, $E\left\{Y^{(1,0)}\right\}$, $E\left\{Y^{(0,1)}\right\}$ and $E\left\{Y^{(1,1)}\right\}$ as follows:

$$\begin{aligned}
\gamma_{\text{int}} &= E\left\{Y^{(0,0)}\right\} = 2.1,\\
\gamma_0 &= E\left\{Y^{(1,0)}\right\} - E\left\{Y^{(0,0)}\right\} = -0.5,\\
\gamma_1 &= E\left\{Y^{(0,1)}\right\} - E\left\{Y^{(0,0)}\right\} = -0.75, \text{and}\\
\gamma_{01} &= E\left\{Y^{(1,1)}\right\} - E\left\{Y^{(1,0)}\right\} - E\left\{Y^{(0,1)}\right\} + E\left\{Y^{(0,0)}\right\} = 0.2.
\end{aligned} \tag{5}$$

Equation (4) is an example of an MSM because it expresses some aspect of the distribution of a potential outcome (in this case, the mean) in terms of the treatment values at that hypothetical intervention and a set of parameters. It is called *marginal* because it is the marginal distribution (unconditional on the time-varying confounder) of the potential outcome that is modelled and *structural* because it is a model for potential outcomes and not a model for the observed outcome. Because A_0 and A_1 are both binary, and the right-hand side of (4) is saturated (there are as many parameters as potential outcomes), there is no parametric smoothing, and thus (4) is a *nonparametric* MSM. In settings with many more timepoints and/or non-binary treatments, MSMs are typically *parametric*, and information from many treatment histories is combined to estimate the parameters of a more parsimonious model (Section 5).

3.4.2. Standard analyses of simulated dataset I. Table I shows the results of a naïve analysis without adjusting for L_1, that is, a regression of Y on A_0, A_1 and A_0A_1 with parameters defined by the equation

$$E(Y|A_0, A_1) = \alpha_{\text{int}} + \alpha_0 A_0 + \alpha_1 A_1 + \alpha_{01} A_0 A_1.$$

Table I also shows the results of a second analysis, adjusting for L_1, that is, a regression of Y on A_0, A_1, A_0A_1 and L_1 with parameters defined by the equation

$$E(Y|A_0, A_1, L_1) = \beta_{\text{int}} + \beta_0 A_0 + \beta_1 A_1 + \beta_{01} A_0 A_1 + \beta_l L_1.$$

Neither analysis is appropriate for drawing causal conclusions, as discussed in Section 3.3. More precisely, neither $(\alpha_0, \alpha_1, \alpha_{01})$ nor $(\beta_0, \beta_1, \beta_{01})$ is, in general, equal to $(\gamma_0, \gamma_1, \gamma_{01})$ as defined by Equation (4).

We provide a demonstration that the standard methods give parameter estimates that tend to limits different from $(\gamma_0, \gamma_1, \gamma_{01})$ (and thus that the biases seen in Table I are not attributable to the finiteness of the sample) in Section B of the Supporting Information.

Table I. The results of the naïve analyses of simulated dataset I, with and without adjusting for L_1, along with the true values of the parameters of (4).

Parameter	Estimate	95% CI	Parameter	Estimate	95% CI	Parameter	True value
α_0	−0.390	(−0.453, −0.327)	β_0	−0.585	(−0.658, −0.511)	γ_0	−0.5
α_1	−0.806	(−0.874, −0.738)	β_1	−0.746	(−0.813, −0.678)	γ_1	−0.75
α_{01}	0.096	(0.002, 0.190)	β_{01}	0.258	(0.160, 0.356)	γ_{01}	0.2

Because the interpretation of the intercept differs according to whether or not we adjust for L_1, we have omitted the intercept (and the coefficient of L_1) from this table.

4. Nonparametric methods

We now turn to two of the three methods proposed by Robins and colleagues, namely the g-computation formula (in Section 4.1) and inverse probability weighted estimation of MSMs (in Section 4.2). Under the assumption of no unmeasured confounding (Section 2.6), these approaches lead to consistent estimators of the causal effect of $\bar{A}$ on Y (Section 2.4), as long as the additional models postulated within each approach are correctly specified. We focus, in this section, on the special case in which these additional models are nonparametric and hence necessarily correctly specified. In this special case, the two methods lead to identical estimates.

4.1. The g-computation formula

4.1.1. The basic idea. In the familiar time-fixed confounding example (Figure 1(a)), a common approach to estimating the expected value of Y^{a_0} in the population (i.e. the expected value of Y under the hypothetical intervention that every member of the population receives treatment a_0) is *standardisation*. This is based on the following formula:

$$E\left(Y^{a_0}\right) = \sum_{l_0 \in \mathcal{L}_0} E\left(Y \,|A_0 = a_0, L_0 = l_0\right) P(L_0 = l_0),$$

where the sum is over all possible values l_0 of L_0. Estimators of $E(Y^{a_0})$ can then be obtained from estimators of $E\left(Y \,|A_0 = a_0, L_0 = l_0\right)$ and $P(L_0 = l_0)$, which can be nonparametric when A_0 and L_0 are discrete.

The *g-computation formula*, introduced by Robins [1], is the appropriate generalisation of standardisation to the setting depicted by Figure 1(c, d), where treatment and covariates vary over time.

4.1.2. A formal description. The g-computation formula is

$$E\left(Y^{\bar{a}}\right) = \sum_{\bar{l} \in \bar{\mathcal{L}}} \left\{ E\left(Y \,|\bar{A} = \bar{a}, \bar{L} = \bar{l}\right) \prod_{t=0}^{T} P\left(L_t = l_t \,|\bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1}\right) \right\}, \tag{6}$$

where the sum is over all possible values $\bar{l}$ of the covariate history.

For the g-computation formula to be implemented, we must therefore estimate $E\left(Y \,|\bar{A} = \bar{a}, \bar{L} = \bar{l}\right)$ and $P\left(L_t = l_t \,|\bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1}\right)$ from the observed data. This involves postulating a model for Y given $\bar{A}$ and $\bar{L}$ and models for L_t given $\bar{A}_{t-1}$ and $\bar{L}_{t-1}$. In the special case where $\bar{A}$ and $\bar{L}$ are binary or categorical, these models can be nonparametric. We fit saturated models so that the mean of Y is separately estimated for each possible combination of treatment and covariate history. Similarly, $P\left(L_t = l_t \,|\bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1}\right)$ can be separately and nonparametrically estimated for each l_t and each combination of $\bar{a}_{t-1}$ and $\bar{l}_{t-1}$.

4.1.3. A demonstration using simulated dataset I. The tree shown in Figure 2 represents the distribution from which dataset I has been generated. The numbers in brackets are the expected numbers of individuals (out of the total of 2000) on each branch of the tree. The numbers inside the circles on the right-hand side are the expected values of Y conditional on being on that branch (i.e. conditional on U_0, A_0, L_1 and A_1). Robins [1] introduced such trees. The branches splitting within a circle depict variables on which we do not (hypothetically) intervene, whereas the branches splitting outside the circles depict the variables of whose causal effects we wish to estimate.

In practice, the observed numbers on each branch are affected by sampling variation, and we do not observe U_0; for now, we ignore these issues.

From the tree in Figure 2, we can construct four trees depicting what we would see if the entire sample were subjected to each of the four hypothetical interventions: $(a_0, a_1) = (0, 0)$, $(a_0, a_1) = (1, 0)$, $(a_0, a_1) = (0, 1)$ and $(a_0, a_1) = (1, 1)$. Notice that in constructing these trees, we force A_0 and A_1 to take particular values, but we allow all the other variables to evolve naturally, according to the probabilities implicit in Figure 2. For example, Figure 3 shows the tree corresponding to $(a_0, a_1) = (0, 0)$. We include the trees corresponding to each of the other three interventions in Section C of the Supporting Information.

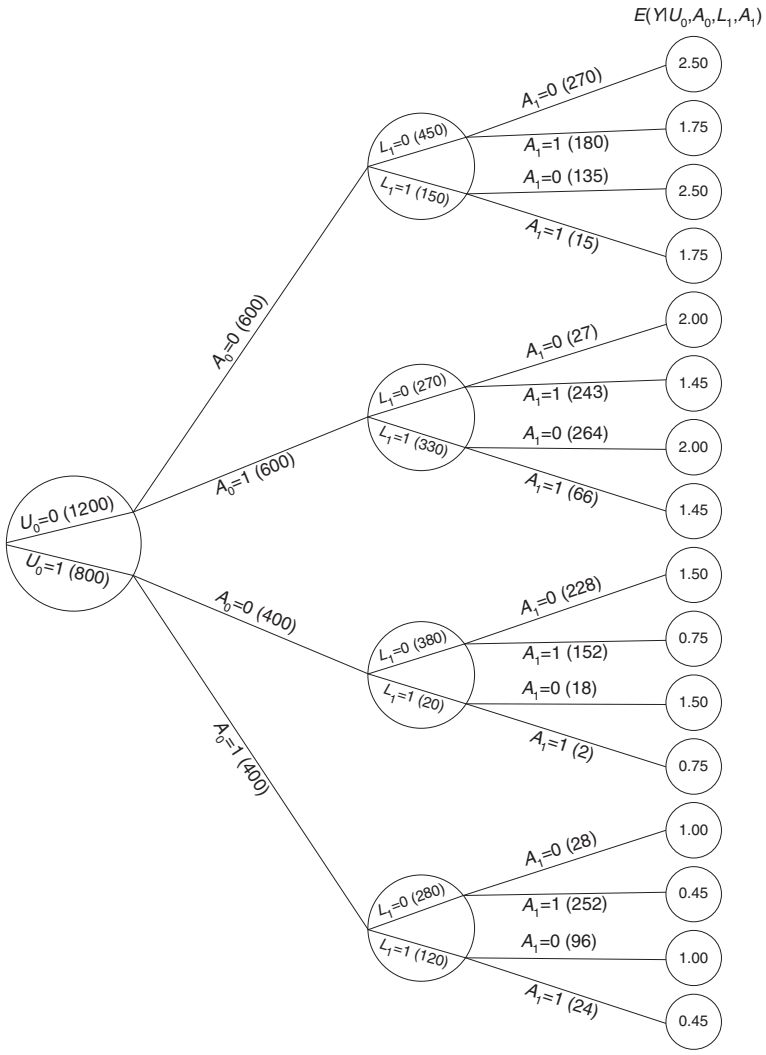

Figure 2. A tree depicting the expected numbers (out of a total of 2000 subjects) along each branch for the distribution from which dataset I has been generated.

Given the four intervention trees, we can calculate $E\left\{Y^{(a_0,a_1)}\right\}$ for each (a_0, a_1). For example, from Figure 3,

$$E\left\{Y^{(0,0)}\right\} = \frac{2.5 \times 900 + 2.5 \times 300 + 1.5 \times 760 + 1.5 \times 40}{2000} = 2.1. \tag{7}$$

What we are doing informally in (7) is applying the formula

$$E\left\{Y^{(0,0)}\right\} = \sum_{u_0,l_1} E(Y \,|U_0 = u_0, A_0 = 0, L_1 = l_1, A_1 = 0)\, P(L_1 = l_1 \,|U_0 = u_0, A_0 = 0)\, P(U_0 = u_0). \tag{8}$$

To understand the intuition behind (8), first write the joint distribution of U_0, A_0, L_1, A_1 (in the observational data) as

$$\begin{aligned} &P(U_0 = u_0)\, P\,(A_0 = a_0 \,|U_0 = u_0)\, P\,(L_1 = l_1 \,|U_0 = u_0, A_0 = a_0) \\ &P(A_1 = a_1 \,|U_0 = u_0, A_0 = a_0, L_1 = l_1)\,. \end{aligned} \tag{9}$$

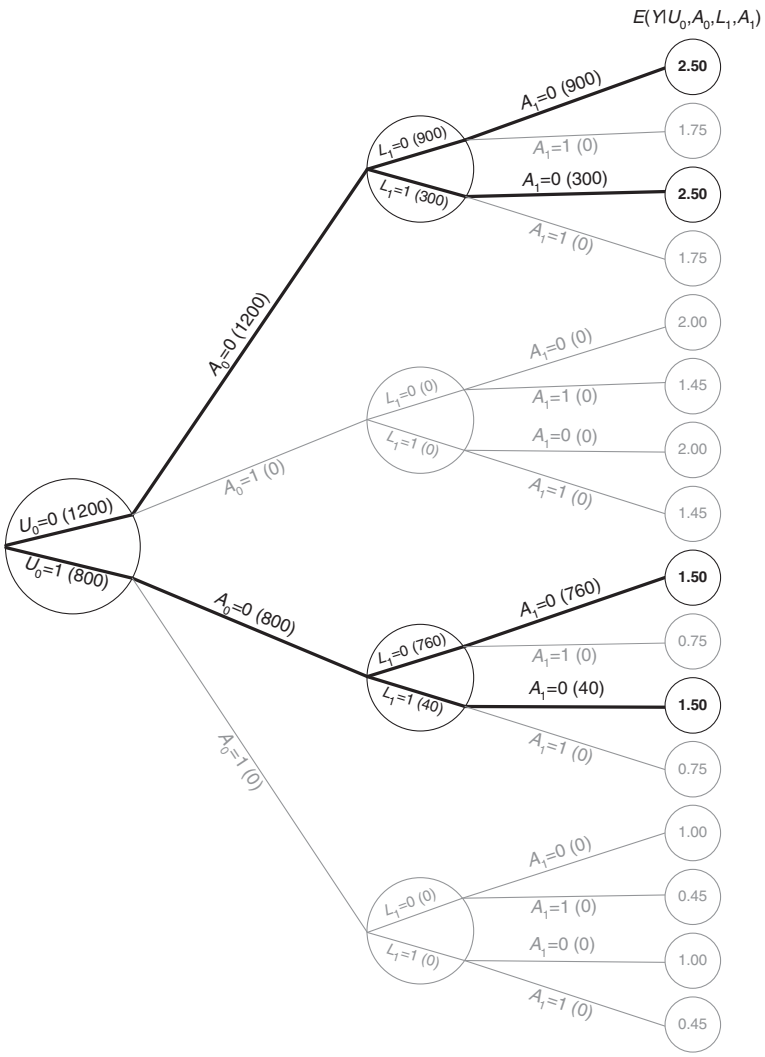

Figure 3. A tree depicting the expected numbers (out of a total of 2000 subjects) along each branch for the distribution from which dataset I has been generated, except that the hypothetical intervention $(a_0, a_1) = (0, 0)$ has been imposed.

$E\left\{Y^{(0,0)}\right\}$ is the mean of Y if, hypothetically, all subjects were forced to have both A_0 and A_1 equal to 0. Under this hypothetical intervention, the distribution of A_0 and A_1 changes so that

$$P(A_0 = 0|U_0 = u_0) = 1, \text{ and}$$
$$P(A_0 = 1|U_0 = u_0) = 0,$$

and

$$P(A_1 = 0|U_0 = u_0, A_0 = 0, L_1 = l_1) = 1, \text{ and}$$
$$P(A_0 = 1|U_0 = u_0, A_0 = 0, L_1 = l_1) = 0.$$

This is equivalent to changing (9) to

$$P(U_0 = u_0)P(L_1 = l_1|U_0 = u_0, A_0 = 0), \tag{10}$$

that is, replacing a_0 and a_1 by 0 throughout and then replacing both $P(A_0 = 0|U_0 = u_0)$ and $P(A_1 = 0|U_0 = u_0, A_0 = 0, L_1 = l_1)$ by 1. Note that the second and third parts of (8) are precisely equal to (10); that is, the formula in (8) standardises the mean of Y to the population from which the sample was drawn under the hypothetical intervention that sets A_0 and A_1 by intervention to 0.

For example, for $U_0 = 0$ and $L_1 = 0$, $P(U_0 = 0)P(L_1 = 1|U_0 = 0, A_0 = 0) = 0.6 \times 0.75 = 0.45 = 900/2000$, the number multiplying $E(Y|U_0 = 0, A_0 = 0, L_1 = 0, A_1 = 0) = 2.2$ in (8).

The formula in (8) specifically estimates the mean of the potential outcome $Y^{(0,0)}$. More generally, we write

$$E\left\{Y^{(a_0,a_1)}\right\} = \sum_{u_0,l_1} E(Y|U_0 = u_0, A_0 = a_0, L_1 = l_1, A_1 = a_1)P(L_1 = l_1|U_0 = u_0, A_0 = a_0)P(U_0 = u_0) \tag{11}$$

for estimating the mean of the potential outcome $Y^{(a_0,a_1)}$.

In practice, we do not observe U_0, and hence we can estimate neither $P(U_0 = u_0)$ from the observed data nor either of the other terms in (11), because they both condition on $U_0 = u_0$. However, in the light of how the counterfactual data were generated, the 'no unmeasured confounding' assumption means that

$$P(A_0 = a_0|U_0 = u_0) = P(A_0 = a_0)$$

and

$$P(A_1 = a_1|U_0 = u_0, A_0 = a_0, L_1 = l_1) = P(A_1 = a_1|A_0 = a_0, L_1 = l_1).$$

Thus, we can re-write (11) using quantities estimable from the observed data alone, as we now demonstrate. We start by re-writing the right-hand side of (11) as follows.[III]

$$\begin{aligned}
&\sum_{u_0,l_1} E(Y|u_0, a_0, l_1, a_1)P(L_1 = l_1|u_0, a_0)P(U_0 = u_0) \\
&= \sum_{l_1}\sum_{u_0} E(Y|u_0, a_0, l_1, a_1)\frac{P(U_0 = u_0)P(A_0 = a_0|u_0)P(L_1 = l_1|u_0, a_0)P(A_1 = a_1|u_0, a_0, l_1)}{P(A_0 = a_0|u_0)P(A_1 = a_1|u_0, a_0, l_1)} \\
&= \sum_{l_1}\sum_{u_0} E(Y|u_0, a_0, l_1, a_1)\frac{P(U_0 = u_0, A_0 = a_0, L_1 = l_1, A_1 = a_1)}{P(A_0 = a_0|u_0)P(A_1 = a_1|u_0, a_0, l_1)} \\
&= \sum_{l_1}\sum_{u_0} E(Y|u_0, a_0, l_1, a_1)\frac{P(A_0 = a_0)P(L_1 = l_1|a_0)P(A_1 = a_1|a_0, l_1)P(U_0 = u_0|a_0, l_1, a_1)}{P(A_0 = a_0|u_0)P(A_1 = a_1|u_0, a_0, l_1)} \\
&= \sum_{l_1}\left\{\sum_{u_0} E(Y|u_0, a_0, l_1, a_1)P(U_0 = u_0|a_0, l_1, a_1)\right\} P(L_1 = l_1|a_0) \\
&\qquad\qquad \text{(by the 'no unmeasured confounding' assumption)} \\
&= \sum_{l_1} E\left(Y|a_0, l_1, a_1\right) P\left(L_1 = l_1|a_0\right). \qquad (12)
\end{aligned}$$

Crucially, (12) does not depend on the distribution of U_0, and thus the procedure carried out on Figure 2 to generate Figure 3 could instead have been performed on the tree shown in Figure 4, where we have averaged over the distribution of $U_0|A_0, L_1, A_1$ to obtain the tree that could be drawn (ignoring random variation) from the observed data.

Thus, we have derived the g-computation formula for our simple example, namely that

$$E\left\{Y^{(a_0,a_1)}\right\} = \sum_{l_1} E(Y|A_0 = a_0, L_1 = l_1, A_1 = a_1)P(L_1 = l_1|A_0 = a_0).$$

Our estimator is then

$$\hat{E}\left\{Y^{(a_0,a_1)}\right\} = \sum_{l_1} \hat{E}(Y|A_0 = a_0, L_1 = l_1, A_1 = a_1)\hat{P}(L_1 = l_1|A_0 = a_0), \tag{13}$$

where $\hat{E}(Y|A_0 = a_0, L_1 = l_1, A_1 = a_1)$ is the empirical average of Y for subjects with $A_0 = a_0$, $L_1 = l_1$, $A_1 = a_1$, and $\hat{P}(L_1 = l_1|A_0 = a_0)$ is the empirical proportion with L_1 equal to l_1

[III] *For ease of notation, we abbreviate* $P(L_1 = l_1|U_0 = u_0, A_0 = a_0)$ *to* $P(L_1 = l_1|u_0, a_0)$ *and so on.*

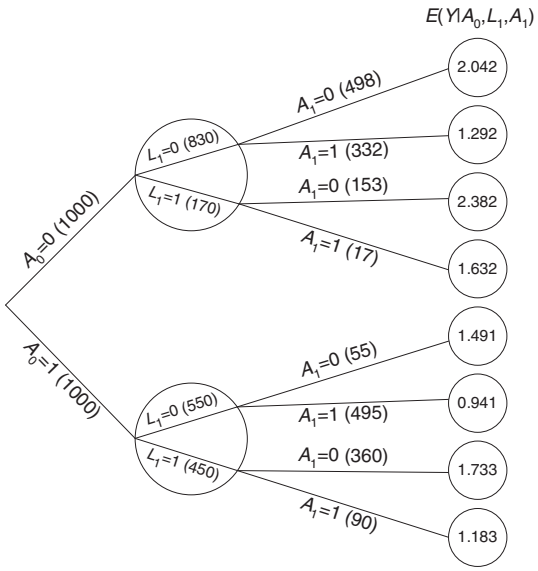

Figure 4. The tree shown in Figure 2, averaged over the distribution of $U_0|A_0, L_1, A_1$.

Table II. The expected values of the four potential outcomes, (a) in truth and (b) as estimated using the g-computation formula on simulated dataset I.

Treatment regime	True mean potential outcome	Estimated mean potential outcome
$a_0 = a_1 = 0$	2.1	2.075
$a_0 = 1, a_1 = 0$	1.6	1.575
$a_0 = 0, a_1 = 1$	1.35	1.336
$a_0 = a_1 = 1$	1.05	1.055

Table III. The results of the analysis of simulated dataset I, as analysed using the g-computation formula to obtain the parameters of the marginal structural model defined in Equation (4), with bootstrap standard errors; the 95% CIs are based on a normal approximation, using the bootstrap standard errors.

Parameter	True value	G-computation estimate	Bootstrap SE	95% CI	
γ_{int}	2.1	2.075	0.021	2.034	2.116
γ_0	−0.5	−0.500	0.047	−0.592	−0.407
γ_1	−0.75	−0.739	0.037	−0.811	−0.667
γ_{01}	0.2	0.218	0.061	0.098	0.339

among subjects with $A_0 = a_0$. Section A2 of the Supporting Information shows the STATA code for this example, and when applied to simulated dataset I leads to the estimates given in Table II. The STATA code given in the Supporting Information is specific to the examples considered and is intended for illustrative purposes only. A more general STATA command is also available [17].

4.1.4. Estimating the parameters of a marginal structural model. We can also obtain the g-computation formula estimates of the parameters of the MSM in Equation (4) using the relationships given in (5). We provide these estimates Table III. In this nonparametric setting, estimating the expected potential outcome under each possible treatment is equivalent to estimating the parameters of the corresponding nonparametric MSM [18]. It is possible in some settings to estimate the standard errors of these estimators using the delta method, but in general, bootstrapping is necessary. The code for this analysis appears in Section A3 of the Supporting Information and the results based on the bootstrap standard errors in Table III.

4.1.5. The need for Monte Carlo methods as T increases. With $T+1$ timepoints (and univariate binary L), there are 2^{T+1} contributions to the sum in (13), and thus implementing the g-computation formula analytically becomes computationally infeasible as T becomes large. The computational burden is even larger for multivariate and/or non-binary L. Robins [1] described a Monte Carlo algorithm to overcome this computational difficulty. The authors in [17, 19] provide more details, and we include a worked example (with STATA code) in Section D of the Supporting Information.

4.1.6. Loss to follow-up. Missing data due to loss to follow-up can be easily incorporated into the g-computation formula, under the assumption that missingness is at random (MAR) [20]. The MAR assumption states that, conditional on staying in the study up to and including visit t, and on $\bar{A}_{t,i}$ and $\bar{L}_{t,i}$, the probability that subject i remains in the study until at least visit $t+1$ is independent of all future variables. Using the g-computation formula, we can view dropping out simply as another possible treatment trajectory, and the hypothetical intervention becomes 'receiving treatment $\bar{a}$ *and* remaining in the study'. We include an example in Section E of the Supporting Information.

4.1.7. Software. The SAS macro GFORMULA performs the aforementioned analysis, allowing for censoring due to death as well as censoring due to loss to follow-up. Monte Carlo simulation is used, as well as bootstrapping for obtaining standard errors. The macro is available from www.hsph.harvard.edu/causal/software.htm; for more details, see [19]. We can also use this macro with parametric models, as described in Section 5.

The authors of this paper have written a similar routine in STATA, which can be downloaded by typing `ssc install gformula` in STATA [17].

4.2. *Inverse probability weighted estimation of marginal structural models*

4.2.1. The basic idea. The basic idea behind inverse probability weighting (IPW) is to re-weight the subjects in the analysis to mimic a situation in which the assignment to treatment is at random. A naïve analysis (ignoring confounders) of the re-weighted sample then leads to consistent estimators of the parameters of the specified MSM (such as (4)).

Alternatively, (stabilised) IPW (Section 4.2.4) can be used to mimic a situation in which the assignment to treatment depends only on baseline covariates. As long as the corresponding MSM is specified conditional on these baseline covariates, the parameters of this MSM can be estimated using a re-weighted version of a standard analysis.

4.2.2. A formal description. The *inverse probability of treatment weight* for a subject i is defined as

$$W_i = \frac{1}{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)}, \tag{14}$$

where $f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(a_t, \bar{a}_{t-1}, \bar{l}_t\right)$ is the conditional probability mass function for A_t given $\left(\bar{A}_{t-1}, \bar{L}_t\right)$ evaluated at $\left(a_t, \bar{a}_{t-1}, \bar{l}_t\right)$, and thus $f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)$ is the same mass function evaluated at the actual values $\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)$ for subject i.[***]

Colloquially, the denominator is thus the probability that the subject receives his or her particular treatment trajectory, conditional on his or her covariate history.

Generalising the informal discussion of Section 3.4.1, we found that an MSM is a model for some aspect of the distribution of the potential outcome associated with treatment trajectory $\bar{a}$, expressed as a known function $h(\cdot)$ of $\bar{a}$ and parameters $\boldsymbol{\gamma}$. For example, when the aspect of interest is the mean, the MSM is written as follows:

$$E\left(Y^{\bar{a}}\right) = h\left(\bar{a}; \boldsymbol{\gamma}\right). \tag{15}$$

More generally, MSMs are models for some aspect of the conditional distribution of the counterfactuals given baseline covariates, as indicated in Section 4.2.4, but are always marginal with respect to post-baseline confounders.

[***] *In probability notation,* $f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)$ *can be written as* $\sum_{a\in\mathcal{A}_t} P\left(A_{t,i}=a \,|\, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right) I(A_{t,i}=a)$.

As discussed in Section 3, structural models concerning potential outcomes, such as (15), are different from associational models, such as

$$E\left(Y \mid \bar{A} = \bar{a}\right) = h\left(\bar{a}; \boldsymbol{\alpha}\right), \tag{16}$$

which are concerned with observed outcomes. The parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ coincide only under the strong condition of *no (measured or unmeasured) confounding*.

We obtain the inverse probability of treatment weighted estimators for the parameters $\boldsymbol{\gamma}$ of the MSM (15) by carrying out the naïve analysis of the observational data (16) after weighting each subject by W_i.

In the re-weighted sample, A_t is independent of $\bar{L}_t$ and $\bar{A}_{t-1}$, and the causal effect of $\bar{A}$ on Y is the same as in the original sample (Section 4.2.5). In the causal diagram of Figure 1(c, d), IPW removes all the arrows *into* $A_0, A_1, \ldots, A_T$ but leaves the arrows *out of* $A_0, A_1, \ldots, A_T$ unchanged. Confounding by $\bar{L}$ has been removed, and thus a naïve analysis of the re-weighted sample leads to a consistent estimator of $\boldsymbol{\gamma}$ as long as the 'no unmeasured confounding' assumption holds and both $h\left(\bar{a}; \boldsymbol{\gamma}\right)$ and the models used to estimate W have been correctly specified.

In order that inverse probability weighted estimation of an MSM be implemented, we must therefore estimate, $f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)$, for each t, from the observed data. This involves postulating, in addition to the MSM, models for A_t given $\bar{A}_{t-1}$ and $\bar{L}_t$. In the special case where $\bar{A}$ and $\bar{L}$ are binary or categorical, these models can be nonparametric.

4.2.3. Stabilised inverse probability weights. If some of the values $f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)$ are close to zero, some of the inverse probability weights above can be very large, leading to considerable imprecision in the estimators of the parameters of the MSM. We obtain an often more stable estimator by re-weighting the sample using the *stabilised inverse probability of treatment weight*, which for subject i is

$$\mathrm{SW}_i = \frac{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1}}\left(A_{t,i}, \bar{A}_{t-1,i}\right)}{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)}.$$

The denominator is the same as for the unstabilised weight, and the numerator is the same as the denominator *without* adjusting for covariate history.

This involves additionally postulating models for A_t given $\bar{A}_{t-1}$. However, misspecifying these models affects the efficiency (and stability) of our estimator but not its consistency.

4.2.4. Stabilised inverse probability weights with the numerator a function of baseline covariates. Sometimes, an MSM conditional on baseline covariates $V \subseteq L_0$ may be of interest:

$$E\left(Y^{\bar{a}} \mid V = v\right) = k\left(\bar{a}, v; \boldsymbol{\eta}\right). \tag{17}$$

This would be the case, for example, if we were interested in effect modification by V.

The stabilised weights can then be modified so that the numerator is conditional on these baseline covariates V [21]:

$$\text{SW-V}_i = \frac{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},V}\left(A_{t,i}, \bar{A}_{t-1,i}, V_i\right)}{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)}. \tag{18}$$

Because V is a subset of L_0, note that the denominator of (18) (as well as the numerator) is conditional on V.

Note also that when using the weights SW-V, there is still confounding by V in the re-weighted sample, and thus an MSM of the form (15) is not appropriate.

4.2.5. A demonstration using simulated dataset I. Let us take the original tree (Figure 2) representing the observed data (including U_0 and ignoring sampling variation) and re-weight using inverse probability weights. With $T = 1$, (14) simplifies to

$$W_i = \frac{1}{f_{A_0}(A_{0,i})\, f_{A_1|A_0,L_1}\left(A_{1,i}, A_{0,i}, L_{1,i}\right)}.$$

Table IV. The values of W for all the combinations of A_0, L_1 and A_1, calculated according to the data generating mechanism for simulated dataset I.

A_0	L_1	A_1	W
0	0	0	10/3
0	0	1	5
0	1	0	20/9
0	1	1	20
1	0	0	20
1	0	1	20/9
1	1	0	5/2
1	1	1	10

For example, from Figure 2,

$$f_{A_1|A_0,L_1}(0,0,0) = P(A_1 = 0|A_0 = 0, L_1 = 0) = \frac{270}{270 + 180} = 0.6,$$

and hence, for subjects with $A_0 = L_1 = A_1 = 0$,

$$W = \frac{1}{0.5 \times 0.6} = \frac{10}{3}.$$

Table IV shows the values of W for all other combinations of A_0, L_1 and A_1.

Next, we apply the weights in Table IV to each branch of the tree in Figure 2. The top branch ($U_0 = A_0 = L_1 = A_1 = 0$) and the ninth branch ($U_0 = 1$, $A_0 = L_1 = A_1 = 0$) are both re-weighted by 10/3. Thus, instead of containing 270 and 228 subjects, they contain 900 and 760, respectively. The other branches are similarly re-weighted, resulting in the tree shown in Figure 5. Notice that A_1 is independent of L_1 in Figure 5, that is, there is no confounding by L_1 in this re-weighted sample: under the assumption of no unmeasured confounding, we have re-constructed data from a randomised experiment, free of confounding, from the confounded observational data.

The code for estimating the parameters of the MSM (4) using IPW appears in Section A4 of the Supporting Information and the results appear in Table V. The parameter estimates agree exactly with those obtained using the nonparametric g-computation formula. There is also very close agreement for the standard errors and confidence intervals. The slight differences are because the standard errors in the weighting approach do not acknowledge the fact that the weights have themselves been estimated from the data. This has been shown in most situations[†††] to lead to overestimation of the standard errors and hence conservative inference [2]. This can be confirmed by comparing Tables III and V, where the standard errors in Table V are slightly larger.

To see why, in this case, both approaches give identical parameter estimates, notice that Figure 5 can be constructed by combining the tree shown in Figure 3 with the other three intervention trees (shown in Section C of the Supporting Information).

4.2.6. Analogy with randomised experiments. Simple (i.e. non-stabilised) IPW aims to create a re-weighted pseudo-sample that mimics random assignment to all the possible interventions, where the probability of being assigned to a particular intervention is the same for all interventions and at all covariate levels.

In stabilised IPW, these probabilities are allowed to be different for different interventions, that is, this mimics a randomised trial in which the treatment groups do not necessarily all contain the same number of subjects. When the numerator of the stabilised weight conditions on baseline covariates, assignment probabilities differ also according to baseline covariates.

The fact that both non-stabilised and stabilised weighting (when all the required assumptions hold) lead to causally interpretable estimators is analogous to the fact that in randomised controlled trials, allocation need not be equal and can depend on variables measured prior to randomisation, as long as these variables are appropriately adjusted for in the analysis.

[†††] *More specifically, in large samples, and when efficient estimators for the parameters of the treatment models are used.*

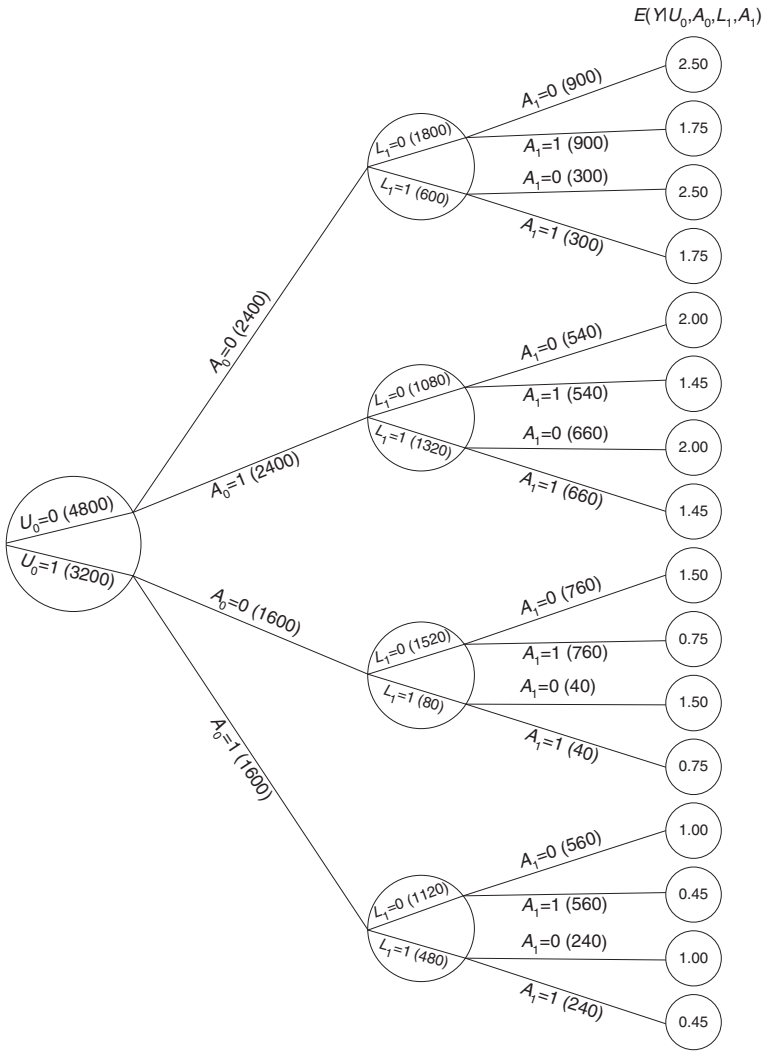

Figure 5. A tree formed by unstabilised inverse probability weighting applied to the distribution from which dataset I has been generated.

Table V. The results of the analysis of the simulated dataset I, as analysed using inverse probability weighting in the marginal structural model (4).

Parameter	True value	IPW estimate	SE‡	95% CI	
γ_{int}	2.1	2.075	0.021	2.034	2.116
γ_0	−0.5	−0.500	0.048	−0.593	−0.406
γ_1	−0.75	−0.739	0.039	−0.815	−0.663
γ_{01}	0.2	0.218	0.065	0.091	0.345

‡This is the sandwich estimator of standard error, which takes into account the non-independence of pseudo-subjects as a result of weighting.

4.2.7. The experimental treatment assignment assumption. IPW requires the *experimental treatment assignment assumption*, namely that

$$0 < f_{A_t|\bar{A}_{t-1},\bar{L}_{t-1}}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t-1,i}\right) < 1 \quad \text{with probability 1} \quad \forall t, i.$$

The probability of receiving the treatment actually received at visit t given the treatment and covariate histories must lie strictly between 0 and 1. To see why this is necessary, consider our analysis of simulated

dataset I. Suppose that all subjects for whom $A_0 = L_1 = 0$ had remained untreated at visit 1, then we would have no information on $E(Y|A_0 = 0, L_1 = 0, A_1 = 1)$, and no amount of re-weighting (indeed, the weight required is infinite) could recover this information. In this case, $E\left\{Y^{(0,1)}\right\}$ would not be estimable.

4.2.8. Loss to follow-up. Write $R_{t,i} = 1$ if subject i is observed at visit t, $R_{t,i} = 0$ otherwise. Then, the missing at random assumption states that

$$P\left(R_{t,i} = 1 \left| \bar{A}_i, \bar{L}_i, R_{t-1,i} = 1\right.\right) = P\left(R_{t,i} = 1 \left| \bar{A}_{t-1,i}, \bar{L}_{t-1,i}, R_{t-1,i} = 1\right.\right).$$

Under this assumption, dealing with loss to follow-up is straightforward.

The stabilised inverse probability weight for subject i at visit t becomes

$$\tilde{W}_i = \frac{\prod_{t=0}^{T}\left\{f_{A_t|\bar{A}_{t-1},R_t}\left(A_{t,i}, \bar{A}_{t-1,i}, 1\right) P\left(R_{t,i} = 1 \left| \bar{A}_{t-1,i}, R_{t-1,i} = 1\right.\right)\right\}}{\prod_{t=0}^{T}\left\{f_{A_t|\bar{A}_{t-1},\bar{L}_t,R_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}, 1\right) P\left(R_{t,i} = 1 \left| \bar{A}_{t-1,i}, \bar{L}_{t-1,i}, R_{t-1,i} = 1\right.\right)\right\}},$$

where $f_{A_t|\bar{A}_{t-1},\bar{L}_t,R_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}, 1\right)$ is the probability mass function for A_t given $\left(\bar{A}_{t-1}, \bar{L}_t, R_t\right)$ evaluated at the actual attained values of $A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}$ and at $R_t = 1$. Colloquially, it is the probability that treatment at visit t takes its attained value given the treatment and confounder histories $\bar{A}_{t-1,i}$, $\bar{L}_{t,i}$ and given that $R_{t,i} = 1$. $f_{A_t|\bar{A}_{t-1},R_t}\left(A_{t,i}, \bar{A}_{t-1,i}, 1\right)$ is similarly defined but without conditioning on $\bar{L}_{t,i}$.

At each visit, the re-weighting can be thought of as operating in two stages: first, the sub-sample remaining in the study is re-weighted to represent multiple copies of itself, one for each of the possible hypothetical interventions, and second, these multiple copies of the sub-sample remaining in the study are re-weighted to represent multiple copies of the original full sample. A worked example (with STATA code) appears in Section F of the Supporting Information.

4.2.9. Software. An advantage of IPW is that bespoke software routines are not, in general, needed, because the models can be fitted using standard regression commands, incorporating weights [22]. However, an MSM routine in SAS is available from www.hsph.harvard.edu/causal/software.htm.

4.3. A comparison of the g-computation formula and inverse probability weighting of marginal structural models

Consider the following factorisation of the (mixed) joint density of all the observed data $O = (A_0, L_1, A_1, Y)$ from our first simulated dataset:

$$f_O(o) = \underbrace{f_{Y|A_0,L_1,A_1}(y, a_0, l_1, a_1)P(L_1 = l_1|A_0 = a_0)}\ \underbrace{P(A_0 = a_0)P(A_1 = a_1|A_0 = a_0, L_1 = l_1)}, \tag{19}$$

where $f_{Y|A_0,L_1,A_1}(y, a_0, l_1, a_1)$ denotes the conditional density function of Y given A_0, L_1, A_1.

We showed earlier (Section 4.1.3) that the mean (or more generally the distribution) of the potential outcome $Y^{(a_0,a_1)}$ can be estimated (by g-computation) using the modified density

$$f_O^*(o) = f_{Y|A_0,L_1,A_1}(y, a_0, l_1, a_1)P(L_1 = l_1|A_0 = a_0) \tag{20}$$

where the second underbrace in (19) has been replaced by 1 (to mimic the hypothetical intervention $(A_0, A_1) = (a_0, a_1)$ applied to the whole sample).

Re-weighting each member of the sample by the inverse of the second underbrace creates a pseudo-sample in which the joint density of $O = (A_0, L_1, A_1, Y)$ is given by $f_O^*(o)$ in (20).

More generally, with T timepoints, the density of the observed data O is factorised as follows:

$$f_O(o) = \underbrace{f_{Y|\bar{A},\bar{L}}\left(y, \bar{a}, \bar{l}\right)\prod_{t=0}^{T} f_{L_t|\bar{A}_{t-1},\bar{L}_{t-1}}\left(l_t, \bar{a}_{t-1}, \bar{l}_{t-1}\right)}\ \underbrace{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(a_t, \bar{a}_{t-1}, \bar{l}_t\right)}. \tag{21}$$

The first underbrace represents the density computed in the g-computation formula, whereas the second underbrace represents that used to define the weights in IPW (if we ignore the stabilising numerator).

This shows the link between the two methods: while the g-computation formula directly calculates the first underbrace and hence the modified distribution of O, $f^*_O(o)$, under any given hypothetical intervention, IPW achieves the same—indirectly—by calculating the second underbrace and re-weighting the data by its inverse. When all models are nonparametric, the two approaches are identical, as was noted by Hernán *et al.* [8] when first introducing IPW of MSMs.

In more realistic settings, potentially misspecified parametric models must be postulated for the first underbrace in the g-computation formula and for the second underbrace in IPW of MSMs. Both sets of parametric models allow the sharing of information between subjects with different treatment and covariate histories, but the extent of this sharing of information (or parametric smoothing) in general differs between the two sets of models. Thus, even if both sets of postulated models are correctly specified (i.e. consistent with the data generating mechanism), we would expect the estimates to differ in finite samples. Under model misspecification, this difference is manifested as a difference in the asymptotic bias of the estimators from the two approaches (IPW and g-computation). In the absence of model misspecification, this difference is manifested as a difference in statistical efficiency. We study the parametric versions of these approaches in the next section.

The parameters directly estimated by the g-computation formula and IPW of MSMs also differ. In the g-computation formula, $E\left\{Y^{(a_0,a_1)}\right\}$ is estimated for each pair of possible values of a_0 and a_1. However, it is the parameters $\boldsymbol{\gamma}$ of the MSM that are estimated by IPW. When each approach is being applied nonparametrically, as in this section, then the two targets of estimation are equivalent. This direct equivalence no longer holds for parametric MSMs, as we see in the next section. Note, however, that the parameters of an MSM can also be estimated using the g-computation formula in a *post hoc* procedure [23]. We describe this in Section 5.1.3.

5. Parametric extensions

5.1. The parametric g-computation formula

The nonparametric g-computation formula easily extends to incorporate parametric models. If all variables in $\bar{L}$ are binary or categorical, then the g-computation formula is still given by Equation (6), the only difference now being that parametric models are postulated for $E\left(Y\left|\bar{A}=\bar{a},\bar{L}=\bar{l}\right.\right)$ and $P\left(L_t=l_t\left|\bar{A}_{t-1}=\bar{a}_{t-1},\bar{L}_{t-1}=\bar{l}_{t-1}\right.\right)$. For continuous L_t, the summation in (6) is replaced by an integral as follows:

$$E\left(Y^{\bar{a}}\right)=\int_{\bar{l}\in\bar{\mathcal{L}}}E\left(Y\left|\bar{A}=\bar{a},\bar{L}=\bar{l}\right.\right)\prod_{t=0}^{T}f_{L_t|\bar{A}_{t-1},\bar{L}_{t-1}}\left(l_t,\bar{a}_{t-1},\bar{l}_{t-1}\right)\mathrm{d}\bar{l},$$

where $f_{L_t|\bar{A}_{t-1},\bar{L}_{t-1}}\left(l_t,\bar{a}_{t-1},\bar{l}_{t-1}\right)$ is the conditional probability density function for L_t conditional on $\bar{A}_{t-1}$ and $\bar{L}_{t-1}$.

This easily extends to multivariate L_t, including the case where L_t contains continuous and discrete variables. An order $L_t=\left(L_t^1,L_t^2\ldots,L_t^{q_t}\right)$ is chosen for the q_t elements of L_t, and then $f_{L_t|\bar{A}_{t-1},\bar{L}_{t-1}}\left(l_t,\bar{a}_{t-1},\bar{l}_{t-1}\right)$ is given by the following:

$$f_{L_t|\bar{A}_{t-1},\bar{L}_{t-1}}\left(l_t,\bar{a}_{t-1},\bar{l}_{t-1}\right)=\prod_{s=1}^{q_t}f_{L_t^s|\bar{A}_{t-1},\bar{L}_{t-1},L_t^1,\ldots,L_t^{s-1}}\left(l_t^s,\bar{a}_{t-1},\bar{l}_{t-1},l_t^1,\ldots,l_t^{s-1}\right).$$

For more details, see [17].

5.1.1. Simulated dataset II. To illustrate the parametric g-computation formula, we introduce a second simulated dataset, where A_0 and A_1 are continuous rather than binary. The causal structure remains that illustrated in Figure 1(e).

Let A_0 and A_1 be the standardised doses of antiglycaemic drug at visits 0 and 1, respectively. Y is standardised log $\mathrm{HbA_{1c}}$ at visit 2; and L_1 is a binary indicator for anaemia at visit 1, measured just before the value of A_1 is chosen.

A sample of 2000 subjects is simulated as follows:

- U_0 is drawn from a Bernoulli distribution with mean 0.4.
- A_0 is drawn from a standard normal distribution.

- Conditional on U_0 and A_0, L_1 is drawn from a Bernoulli distribution with

$$P(L_1 = 1|A_0, U_0) = \frac{\exp(-1.2 + 0.5A_0 - U_0)}{1 + \exp(-1.2 + 0.5A_0 - U_0)}.$$

  That is, anaemia at visit 1 is more common among unhealthy subjects, and the higher the dose of treatment received at visit 0, the higher the probability of anaemia at visit 1. In our simulated dataset, 365 of these 2000 subjects are anaemic at visit 1 ($L_1 = 1$).
- Conditional on A_0 and L_1, A_1 is drawn from a normal distribution with mean $0.2 + 0.6A_0 - L_1$ and standard deviation 0.75, that is, being anaemic at visit 1 decreases the average dose at visit 1, and a high dose at visit 0 is associated with a high dose at visit 1.
- Finally, conditional on U_0, A_0 and A_1, standardised log HbA_{1c} (Y) is generated from a normal distribution with mean $0.2 - 0.12A_0 - 0.2A_1 - 0.5U_0$ and standard deviation 0.93. There is an increased beneficial causal effect of treatment the greater the dose, and this increase is linear. Conditional on treatment, HbA_{1c} is also higher on average for the unhealthy patients.
- Because the aforementioned description is of the data generating process, the parameters are all structural (i.e. causal). In other words, the counterfactual means are given by averaging over the distribution of U_0:

$$E\left\{Y^{(a_0,a_1)}\right\} = 0.2 - 0.12a_0 - 0.2a_1 - 0.5 \times 0.4 = -0.12a_0 - 0.2a_1. \quad (22)$$

  Equivalently, the potential outcomes $Y_i^{(a_0,a_1)}$ can be thought of as being generated independently for each (a_0, a_1) and for each i from a normal distribution with mean $0.2 - 0.12a_0 - 0.2a_1 - 0.5U_{0,i}$ and standard deviation 0.93.

Following the same argument given in Section 3.4.1, $A_0 \perp\!\!\!\perp \left\{Y^{(a_0,a_1)}\right\}$ and $A_1 \perp\!\!\!\perp \left\{Y^{(a_0,a_1)}\right\}|A_0, L_1$, that is, the data have been generated under the 'no unmeasured confounding' assumption.

The STATA code used to simulate this dataset appears in Section A5 of the Supporting Information, and the dataset is available from the corresponding author upon request.

5.1.2. Analysing simulated dataset II using the parametric g-computation formula. We specify the following parametric models:

$$E(Y|A_0, L_1, A_1) = \mu + \nu_0 A_0 + \nu_1 A_1 + \nu_{01} A_0 A_1 + \nu_l L_1 + \nu_{0l} A_0 L_1 + \nu_{1l} A_1 L_1 + \nu_{01l} A_0 A_1 L_1 \quad (23)$$

and

$$P(L_1 = 1|A_0) = \frac{\exp\left(\zeta + \eta_0 A_0 + \eta_{02} A_0^2\right)}{1 + \exp\left(\zeta + \eta_0 A_0 + \eta_{02} A_0^2\right)}.$$

These models are fitted to the observed data and the fitted values $\hat{E}(Y|A_0, L_1, A_1)$ and $\hat{P}(L_1 = 1|A_0)$ substituted into (13) for any pair of values (a_0, a_1) to estimate $E\left\{Y^{(a_0,a_1)}\right\}$.

The STATA code given in Section A6 of the Supporting Information calculates $E\left\{Y^{(a_0,a_1)}\right\}$ for specified values of a_0 and a_1.

5.1.3. Estimating the parameters of an MSM. It is possible to use our estimates of $E\left\{Y^{(a_0,a_1)}\right\}$ (for a range of values of a_0 and a_1) to obtain *post hoc* estimates of the parameters of an MSM [23]. Specifically, we consider again the following MSM:

$$E\left\{Y^{(a_0,a_1)}\right\} = \gamma_{\text{int}} + \gamma_0 a_0 + \gamma_1 a_1 + \gamma_{01} a_0 a_1, \quad (24)$$

which (in contrast to the nonparametric setting with binary A_0 and A_1) is no longer necessarily correctly specified (although, in fact, given our knowledge of the data generating process, we know that it is). As well as being potentially misspecified, in general, it is possible for the MSM to be incompatible with the parametric associational models specified in the g-computation formula. For example, if (24) did not include the treatment-by-treatment interaction term and the associational model (23) did, or vice versa,

Table VI. The results of the analysis of simulated dataset II, using the g-computation formula (with bootstrap standard errors) to estimate the parameters of the marginal structural model (24).

Parameter	True value	G-computation estimate	Bootstrap SE	95% CI	
γ_{int}	0	−0.0255	0.0265	−0.0775	0.0265
γ_0	−0.12	−0.1260	0.0274	−0.1797	−0.0722
γ_1	−0.2	−0.2242	0.0309	−0.2847	−0.1637
γ_{01}	0	0.0063	0.0207	−0.0343	0.0469

then they would be incompatible with each other. Furthermore, if Y were binary, then logistic regression models for both (23) and (24) would be incompatible with each other as a result of non-collapsibility.

For each subject i, we calculate $E\left\{Y^{(A_{0,i},A_{1,i})}\right\}$, the expected potential outcome associated with the observed treatment levels for this subject. Then, we fit a linear regression model with $E\left\{Y^{(A_{0,i},A_{1,i})}\right\}$ as the outcome variable and $A_{0,i}$, $A_{1,i}$ and $A_{0,i}A_{1,i}$ as explanatory variables. This gives estimators of γ_{int}, γ_0, γ_1 and γ_{01}, and we use the bootstrap to obtain their standard errors. This code also appears in Section A6 of the Supporting Information and the results appear in Table VI.

5.1.4. The g-null paradox. In addition to the assumption of no unmeasured confounding, the consistency of our estimator of $E\left\{Y^{(a_0,a_1)}\right\}$ now relies on having correctly specified the parametric forms of $E(Y|A_0, L_1, A_1)$ and $P(L_1 = l_1|A_0)$.

Robins in his original paper [1] highlighted a problem with the parametric version of the g-computation formula described above, and Robins and Wasserman [15] elaborated further on this. When standard, parsimonious models are chosen for the conditional distributions of Y given $\left(\bar{A}, \bar{L}\right)$ and L_t given $\left(\bar{A}_{t-1}, \bar{L}_{t-1}\right)$, then, when the causal null hypothesis is true, upon combining them using the g-computation formula, the implied distribution for $Y^{\bar{a}}$ is typically given by a complex function of $\bar{a}$, such that the distribution of $Y^{\bar{a}}$ is independent of $\bar{a}$ only under the additional constraint that L_t is not affected by $\bar{A}_{t-1}$ (i.e. when standard methods would in any case suffice).

This is known as the *g-null paradox*: given enough data, we would reject the causal null hypothesis even when it is true.

For example, suppose that L_1 is binary and A_0 is continuous, then (as was carried out previously) it is natural to model L_1 conditional on A_0 using a logistic regression model, but Robins and Wasserman showed that, whenever A_0 has a non-null effect on L_1, this model is inconsistent with the causal null hypothesis. The g-null paradox is of particular concern when interest lies in testing the causal null hypothesis, and Robins and Hernán [4] recommended that use of the parametric g-computation formula be avoided whenever the causal null hypothesis cannot be ruled out.

5.1.5. Model misspecification in the analysis of simulated dataset II. In simulated dataset II, L_1 was generated using a non-trivial logistic function of A_0 and U. Because of the non-collapsibility of the odds ratio [24, 25], it is not possible for $P(L_1 = 1|A_0)$ to be a logistic function of A_0 after marginalising over U, and thus our analysis model is misspecified. We see in Figure 6 (similar results were seen for the other three parameters) that the resulting bias is virtually undetectable in this example. Notice that in our logistic regression model for L_1, we in fact included A_0^2 as a covariate in addition to A_0. The use of higher-order terms in this way somewhat reduces the impact of misspecification.

5.2. *Inverse probability weighted estimation of marginal structural models: parametric setting*

When L is continuous (but not A), inverse probability weighted estimation of MSMs works similarly to the nonparametric setting. The only difference is that models for A_t given $\bar{A}_{t-1}$ and $\bar{L}_t$ must now be specified parametrically (e.g. a logistic regression if A is binary).

If A is continuous, then, in principle, the approach can still be applied with the mass functions used to define the weights replaced by densities, although these typically suffer from greater inefficiency, instability and greater risk of model misspecification than in the discrete case, as is discussed later.

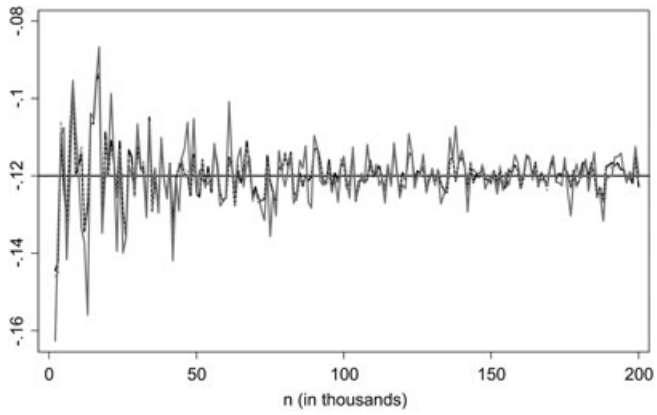

Figure 6. A comparison of the parameter estimates for the coefficient of a_0 from the g-computation formula (black) and IPW (grey) analyses of datasets simulated according to the same distribution as simulated dataset II, along with the true parameter value, as the sample size increases. The estimates of ϕ_4 obtained from g-estimation (discussed later) are also comparable and hence shown (using a dotted line).

Estimators based on the stabilised weights

$$\text{SW}_i = \frac{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1}}\left(A_{t,i}, \bar{A}_{t-1,i}\right)}{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)}$$

or

$$\text{SW-V}_i = \frac{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},V}\left(A_{t,i}, \bar{A}_{t-1,i}, V_i\right)}{\prod_{t=0}^{T} f_{A_t|\bar{A}_{t-1},\bar{L}_t}\left(A_{t,i}, \bar{A}_{t-1,i}, \bar{L}_{t,i}\right)}$$

should be used and an appropriate MSM fitted to the re-weighted sample as discussed in Section 4.2.

5.2.1. Analysing simulated dataset II using inverse probability weighted estimation of the marginal structural model. A normal density is used for both $A_1|A_0$ and $A_1|A_0, L_1$, with the conditional mean modelled as

$$E(A_1|A_0, L_1) = \kappa + \xi_0 A_0 + \xi_l L_1 + \xi_{0l} A_0 L_1, \tag{25}$$

in accordance with the way in which the data were generated.

It is important to note that this approach requires the correct specification of the density (in this case normal) as well as the conditional expectation in (25), and even mild misspecifications of the density function can lead to severe biases.

The code for this analysis appears in Section A7 of the Supporting Information, and Table VII shows the results.

Table VII. The results of the analysis of simulated dataset II, using inverse probability weighted estimation of the marginal structural model (24).

Parameter	True value	IPW estimate	SE	95% CI	
γ_{int}	0	−0.0183	0.0299	−0.0770	0.0405
γ_0	−0.12	−0.1104	0.0309	−0.1710	−0.0499
γ_1	−0.2	−0.2189	0.0462	−0.3096	−0.1282
γ_{01}	0	0.0179	0.0281	−0.0372	0.0731

5.3. Comparing g-computation and inverse probability weighting in the parametric setting

Tables VI and VII show that the results from the two approaches are similar, but not identical, with slightly larger standard errors from IPW compared with g-computation. We show a further comparison of the estimates of the coefficient of a_0 (in Equation (24)) in Figure 6. We will discuss the comparison with g-estimation in the next section. We generated a series of datasets with the same distribution as simulated dataset II, but with increasing sample size. The graph shows how the estimates change as the sample size increases. We also show the true parameter value. We see that IPW is indeed somewhat less efficient and less stable than the g-computation formula, although the difference is not striking here. The inefficiency and instability of IPW increases as the dimension of L increases and is a common phenomenon discussed in the literature; see for example [26]. As the dimension of L increases, the inverse probability weights (even after stabilisation) typically increase in variability, giving relatively very large weights to very few subjects in the dataset.

The difference between estimation methods seen in Figure 6 is in contrast to our observation in Section 4, where the g-computation formula and IPW gave identical results. Once modelling becomes non-trivial, it is usually more powerful to model directly the terms included in the first underbrace of Equation (21) than to get at these indirectly by factorising out the terms included in the second underbrace [26].

The gain in efficiency from using the g-computation formula over IPW comes at the cost, however, of increased risk of bias due to model misspecification. When a particular subject i is given a large weight in the analysis, it is indicative of the fact that there were not many subjects in the study who behaved similarly to i and that therefore, to make inference about what might happen to the population if they all behaved similarly to subject i, we must rely heavily on the particular outcome of subject i, which is subject to random variation. The g-computation formula would, in this situation, fit a global model to Y conditional on past treatment and covariate histories, and subject i would lie far away from the covariate 'centre' of the data used to fit this model. Any predictions made for the whole population under behaviour similar to that of subject i's would likely be based on extrapolation beyond the region for which the model fit could be reliably assessed. This is an example of the familiar trade-off between statistical efficiency and strength of assumptions. In our example, the model for Y conditional on past treatment and covariate histories used in the g-computation formula analysis is entirely consistent with the model used to generate the data, and thus the gamble pays off and we see efficiency gains without introducing bias [27].

6. G-estimation of structural nested models

Robins [5, 28, 29] first introduced g-estimation of SNMs, and this is closely related to the g-test of the causal null hypothesis, which we introduce in Section 6.1. In Section 6.2, we give a formal description of the method before using it to analyse simulated dataset II in Section 6.3. We start with the parametric setting as it is the most natural way of introducing g-estimation; however, in Section 6.5, we return to re-analyse simulated dataset I to show that the equivalence of these approaches in the nonparametric setting extends to g-estimation.

6.1. The g-test of the causal null hypothesis

Suppose we wanted to test the hypothesis that A_0 and A_1 have no causal effect on Y in our simplified setting illustrated by Figure 1(e), that is, that $E\left\{Y^{(a_0,a_1)}\right\}$ is the same for all (a_0, a_1). We assume throughout that the 'no unmeasured confounding' assumption holds, that is, that

$$A_0 \perp\!\!\!\perp \left\{Y^{(a_0,a_1)}\right\}$$

and

$$A_1 \perp\!\!\!\perp \left\{Y^{(a_0,a_1)}\right\} \Big| A_0, L_1.$$

Usually, we make the point that the 'no unmeasured confounding' assumption cannot be tested, because at most one of the set $\left\{Y^{(a_0,a_1)}\right\}$ is observed on each subject. However, under the causal null hypothesis, all of the $Y^{(a_0,a_1)}$ are equal and therefore equal to the observed outcome Y. Thus, under the null hypothesis, the 'no unmeasured confounding' assumption becomes

$$A_0 \perp\!\!\!\perp Y$$

and

$$A_1 \perp\!\!\!\perp Y | A_0, L_1.$$

Thus, we can test the causal null hypothesis and the 'no unmeasured confounding' assumption together by postulating models for A_0 and $A_1|A_0, L_1$, including Y as an additional covariate in these models, and then testing (more details in Section 6.3) the hypothesis that the coefficient of Y is zero in both models. A small p-value provides evidence against the combined assumptions of 'no unmeasured confounding' and the causal null hypothesis. Then, on the basis of the *assumption* of no unmeasured confounding, evidence against this combination becomes evidence against the causal null hypothesis.

This is the logic behind the g-test of the causal null hypothesis. For each t, we postulate a model for A_t conditional on $\bar{A}_{t-1}$ and $\bar{L}_t$. We fit these models, adding Y as an additional covariate in each. If there is evidence to reject the hypothesis that the coefficient of Y is zero in all of these models, this is evidence against the causal null hypothesis. Note that, as was the case in IPW, this involves specifying models for each A_t given $\bar{A}_{t-1}$ and $\bar{L}_t$, without needing to specify models for each L_t given $\bar{A}_{t-1}$ and $\bar{L}_{t-1}$, or for Y given $\bar{A}$ and $\bar{L}$.

6.2. *A formal description of structural nested models and g-estimation*

6.2.1. The basic idea. G-estimation is the inversion of the g-test [30, p. 420]. We search for a causal effect under which the set of potential outcomes is conditionally independent of the time-varying treatment given the histories of the time-varying treatment and confounder. For the inversion of the g-test to be possible, we must first characterise the causal effect of interest in terms of a finite number of parameters, that is, using a structural model. Because of the nature of the estimation technique, it turns out that a different sort of structural model is required from the MSM introduced earlier. We now introduce this class of models, known as *SNMs*.

6.2.2. Structural nested model. An SNM consists of a set of sub-models, one for each of $t = 0, \ldots, T$, where in the tth sub-model, a relationship is postulated between some aspect of the laws of $Y_i^{(a_0,a_1,\ldots,a_{t-1},a_t,0,\ldots,0)}$ and $Y_i^{(a_0,a_1,\ldots,a_{t-1},0,0,\ldots,0)}$. The tth sub-model is thus a comparison of two potential outcomes. The first potential outcome would have been observed on subject i had he or she followed a particular treatment trajectory $\bar{a}$ up to and including timepoint t and no treatment at all thereafter. The second would have been observed had he or she followed the same treatment trajectory $\bar{a}$ but only up to and including timepoint $t-1$ and no treatment at all at visit t or thereafter. The tth sub-model thus characterises the effect of a final amount a_t of treatment at visit t. In general, the sub-model for $Y_i^{(a_0,a_1,\ldots,a_{t-1},a_t,0,\ldots,0)}$ in terms of $Y_i^{(a_0,a_1,\ldots,a_{t-1},0,0,\ldots,0)}$ and a_t is expressed as a function of $\bar{a}_{t-1}$ and $\bar{l}_t$ for the subset of subjects with $\bar{A}_{t-1,i} = \bar{a}_{t-1}$ and $\bar{L}_{t,i} = \bar{l}_t$. In this general formulation, the effect of a final amount a_t of treatment at visit t is allowed to differ according to past values of the treatment and covariates. This is the key difference between marginal structural and SNMs: the former are marginal with respect to the time-changing confounders, whereas the latter are conditional on them. The $T+1$ sub-models making up the SNM are defined in terms of parameters $\boldsymbol{\phi}$. These are estimated using g-estimation, as we describe in Section 6.2.5. The exact way in which the $Y_i^{(a_0,a_1,\ldots,a_{t-1},a_t,0,\ldots,0)}$ and $Y_i^{(a_0,a_1,\ldots,a_{t-1},0,0,\ldots,0)}$ are related as a function of $\boldsymbol{\phi}$ and treatment/covariate histories depends on the type of outcome Y (binary, continuous, etc.), and thus, to give more detail, we turn to one particular example, namely the structural nested mean model (SNMM), which is used when Y is continuous.

6.2.3. Structural nested mean model. Robins [29] proposed the SNMM for continuous outcomes.[‡‡‡] For each $t \in \{0, 1, \ldots, T\}$, the SNMM is given as

$$\begin{aligned} &E\left\{Y^{(a_0,a_1,\ldots,a_{t-1},a_t,0,\ldots,0)} \,\middle|\, \bar{A}_{t-1,i} = \bar{a}_{t-1}, \bar{L}_{t,i} = \bar{l}_t\right\} \\ &\quad = E\left\{Y^{(a_0,a_1,\ldots,a_{t-1},0,0,\ldots,0)} \,\middle|\, \bar{A}_{t-1,i} = \bar{a}_{t-1}, \bar{L}_{t,i} = \bar{l}_t\right\} + \psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right). \end{aligned} \tag{26}$$

[‡‡‡] *Robins has also proposed the structural nested failure time model* [3] *for time-to-event outcomes, but we do not discuss this further here. See* [6,7] *for further details. The SURV macro in* SAS *(available from www.hsph.harvard.edu/causal/software.htm) and the stgest command in* STATA [31] *both implement this method.*

The function $\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right)$ is called the tth *blip function*, because it characterises the effect of a final 'blip' of treatment a_t at visit t. This function is specified in terms of a_t and the treatment and covariate histories up to visit t, using a set of parameters $\boldsymbol{\phi}_t$. For example, a simple blip function is

$$\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right) = \phi_{t,0} a_t. \tag{27}$$

This assumes that the effect of the final blip a_t of treatment is linear in a_t and the same regardless of treatment and covariate history. An alternative blip function is given by

$$\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right) = (\phi_{t,0} + \phi_{t,1} a_{t-1} + \phi_{t,2} l_t) a_t. \tag{28}$$

This allows the effect of the final blip a_t of treatment to differ according to previous treatment and current covariate.

A blip function $\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right)$ must satisfy the following condition:

$$\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right) = 0 \begin{cases} \text{if } a_t = 0 \\ \text{if } \boldsymbol{\phi}_t = \mathbf{0}. \end{cases}$$

The first condition is necessary for (26) to hold at $a_t = 0$, and the second condition means that $\boldsymbol{\phi} = \mathbf{0}$ corresponds to the causal null hypothesis of no causal effect of treatment.

6.2.4. The 'blipped down' and 'treatment-free' counterfactuals. We write $\boldsymbol{\phi} = (\boldsymbol{\phi}_0, \ldots, \boldsymbol{\phi}_T)$ and define

$$Y^*_{t,i}(\boldsymbol{\phi}) = Y_i - \sum_{s=t}^{T} \psi_s\left(\bar{A}_{s,i}, \bar{L}_{s,i}; \boldsymbol{\phi}_s\right)$$

to be the 'stage t blipped down' version of Y_i. Under the true values of $\boldsymbol{\phi}$, and assuming correct specification of the SNMM (i.e. correct specification of the blip functions), the expected value of $Y^*_{t,i}(\boldsymbol{\phi})$ is the same as the expected value of Y_i had subject i received her actual observed treatment trajectory up to and including timepoint $t-1$ and no treatment at all thereafter.

$$Y^*_{0,i}(\boldsymbol{\phi}) = Y_i - \sum_{s=0}^{T} \psi_s\left(\bar{A}_{s,i}, \bar{L}_{s,i}; \boldsymbol{\phi}_s\right)$$

is called the 'treatment-free counterfactual'. Assuming correct specification of the SNMM, and under the true values of $\boldsymbol{\phi}$, the expected value of $Y^*_{0,i}$ is the expected value of Y had subject i remained untreated for the entire follow-up.

6.2.5. G-estimation of the parameters of a structural nested model. Conceptually, the parameters $\boldsymbol{\phi}$ could be estimated using a searching algorithm as follows.

First, for each $t = 0, \ldots, T$, a model for A_t conditional on $\bar{L}_t$ and $\bar{A}_{t-1}$ would be postulated. Then, for a particular set of values of $\boldsymbol{\phi}$, $Y_t^*(\boldsymbol{\phi})$ would be calculated and added as an additional covariate to this model. A joint significance test would then be performed to test the hypothesis that the coefficient of $Y_t^*(\boldsymbol{\phi})$ in the model for A_t is equal to zero, for each $t = 0, \ldots, T$. If the p-value from this test were greater than 0.05, then this particular set of values of $\boldsymbol{\phi}$ would be included in the 95% confidence region for $\boldsymbol{\phi}$. This procedure would be repeated for many values of $\boldsymbol{\phi}$ to construct the confidence region. The values of $\boldsymbol{\phi}$ corresponding to a p-value of 1 would be our g-estimates of $\boldsymbol{\phi}$.

When $\boldsymbol{\phi}$ is d-dimensional for $d > 2$, this sort of searching algorithm is computationally infeasible.

However, whenever $\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right)$ is of the form

$$\psi_t\left(\bar{a}_t, \bar{l}_t; \boldsymbol{\phi}_t\right) = \boldsymbol{\phi}' a_t v_t\left(\bar{a}_{t-1}, \bar{l}_t\right), \tag{29}$$

where $v_t\left(\bar{a}_{t-1}, \bar{l}_t\right)$ is a $(d \times 1)$ vector of functions of $\bar{a}_{t-1}$ and $\bar{l}_t$, and $'$ denotes matrix transpose, a closed-form solution for the g-estimates of $\boldsymbol{\phi}$ exists [4, 29] as we now describe. Note that the left-hand side of (29) is a function of $\boldsymbol{\phi}_t$, whereas the right-hand side is a function of $\boldsymbol{\phi}$. In order that the equality hold, the entries of $v_t\left(\bar{a}_{t-1}, \bar{l}_t\right)$ corresponding to elements of $\boldsymbol{\phi}$ not in $\boldsymbol{\phi}_t$ are set to zero. For example, the blip function given in (27) is associated with the set of $(T \times 1)$ vectors $v_t\left(\bar{a}_{t-1}, \bar{l}_t\right)$ that have a one at the tth entry and zeros elsewhere. The blip function given in (28) is associated with the set of $(3T \times 1)$

vectors $v_t\left(\bar{a}_{t-1}, \bar{l}_t\right)$ that have the first $3(t-1)$ entries equal to zero, followed by $(1, a_{t-1}, l_t)'$, followed by a further $3(T-t)$ zeros.

Essentially, the following estimating equation is solved for $\boldsymbol{\phi}$:

$$\sum_{i=1}^{n}\sum_{t=0}^{T}\left(A_{t,i}-\hat{A}_{t,i}\right)B_{t,i}Y_{t,i}^{*}(\boldsymbol{\phi})=0$$

where $\hat{A}_{t,i}$ is the predicted value of $A_{t,i}$ from the model for A_t given $\left(\bar{A}_{t-1}, \bar{L}_{t-1}\right)$ fitted to the observational data and $B_{t,i}$ is a d-dimensional vector of functions of $\bar{A}_{t-1,i}$ and $\bar{L}_{t,i}$. Intuitively, we see that we are essentially setting the conditional covariance of Y_t^* and A_t given $\bar{A}_{t-1}$ and $\bar{L}_t$ to zero for each t, exactly as described in the aforementioned searching algorithm.

The choice of $B_{t,i}$ affects efficiency but not consistency, and the optimally efficient choice has been described [29]. A simple and reasonable choice of $B_{t,i}$ is to take $B_{t,i} = v_t\left(\bar{A}_{t-1,i}, \bar{L}_{t,i}\right)$ [4]. Further efficiency gains may be achieved by subtracting from $Y_{t,i}^{*}(\boldsymbol{\phi})$ its conditional expectation.

Then, the g-estimate of $\boldsymbol{\phi}$ is given by

$$\hat{\boldsymbol{\phi}}=\left\{\sum_{i=1}^{n}\sum_{t=0}^{T}\left(A_{t,i}-\hat{A}_{t,i}\right)B_{t,i}\sum_{k=t}^{T}A_{k,i}B_{k,i}'\right\}^{-1}\left\{\sum_{i=1}^{n}\sum_{t=0}^{T}\left(A_{t,i}-\hat{A}_{t,i}\right)B_{t,i}Y_i\right\}. \tag{30}$$

We illustrate this using an example in Section 6.3.

6.2.6. Why is it called a structural nested model? At each timepoint t, we consider a (hypothetical) population, let us call it population 1, consisting of subjects who have, up to visit t, experienced identical histories $\left(\bar{a}_{t-1}, \bar{l}_t\right)$. Then, assuming that no one in population 1 is to be treated from visit $t+1$ onwards, we ask what would be the causal effect on Y of a final amount a_t of treatment at visit t in population 1. Then, we would move on to visit $t+1$ and ask what would be the causal effect on Y of a final amount a_{t+1} of treatment at visit t in a (hypothetical) population 2, consisting of subjects who have, up to visit $t+1$, experienced identical histories $\left(\bar{a}_t, \bar{l}_{t+1}\right)$. Population 2 is a subset of population 1: this explains the term *nested* in SNM. As for MSMs, the term *structural* refers to the fact that it is a model relating potential, rather than observed, outcomes.

6.3. Analysing simulated dataset II using g-estimation of a structural nested mean model

We consider the following SNMM for simulated dataset II:

$$\psi_1(a_0, a_1, l_1; \boldsymbol{\phi}) = (\phi_0 + \phi_1 a_0 + \phi_2 l_1 + \phi_3 a_0 l_1)a_1 \tag{31}$$

and

$$\psi_0(a_0; \boldsymbol{\phi}) = \phi_4 a_0. \tag{32}$$

First, we fit the model shown in (25) and obtain predictions

$$\hat{A}_{1,i} = \hat{\kappa} + \hat{\xi}_0 A_{0,i} + \hat{\xi}_l L_{1,i} + \hat{\xi}_{0l}\hat{A}_0\hat{L}_1.$$

For A_0, the prediction for each i is simply the sample mean

$$\hat{A}_{0,i} = \frac{1}{2000}\sum_{j=1}^{2000}A_{0,j}.$$

We take

$$B_{1,i} = (0, 1, A_{0,i}, L_{1,i}, A_{0,i}L_{1,i})'$$

and

$$B_{0,i} = (1, 0, 0, 0, 0)'.$$

We then plug these quantities into (30) to obtain the g-estimates of $\boldsymbol{\phi}$ and obtain standard errors using a sandwich estimator (see Section G of the Supporting Information for further details). Section A8 of the Supporting Information displays the STATA code for this analysis, and Table VIII shows the results.

Table VIII. The results of the analysis of simulated dataset II, using g-estimation of the structural nested model given by blip functions (31) and (32).

Parameter	True value	g-estimate	SE	95% CI	
ϕ_0	−0.2	−0.2287	0.0326	−0.2925	−0.1649
ϕ_1	0	−0.1278	0.0318	−0.0750	0.0495
ϕ_2	0	−0.1904	0.0836	−0.1829	0.1448
ϕ_3	0	0.0991	0.0850	−0.0674	0.2657
ϕ_4	−0.12	−0.1336	0.0265	−0.1856	−0.0815

6.4. Comparing g-estimation with g-computation and inverse probability weighting

Although the parameters of the SNM defined by (31) and (32) do not in general correspond to the parameters of an MSM, because of the (interaction-free) way in which the data were generated, in this case, we do have correspondence between two of these parameters and those of the MSM (24), namely $\phi_4 = \gamma_0$ and $\phi_0 = \gamma_1$. Looking only at these parameters and comparing Tables VI–VIII and comparing the g-estimates of ϕ_4 with the estimates of γ_0 shown in Figure 6, we find that the results are similar from all three methods, with the efficiency of g-estimation being similar to g-computation and better than IPW. In general, as the dimension of L increases, the differences in efficiency become more stark, and g-estimation typically lies between g-computation and IPW in terms of efficiency. The increase in efficiency of g-estimation compared with IPW is interesting; the associational models specified in each are exactly the same, namely the treatment process conditional on covariate and treatment histories is modelled, but the remaining parts of the joint likelihood of the observed data are left unspecified. The difference in efficiency between g-estimation and IPW is rather a consequence of how these models are used to estimate causal parameters from different structural models; the assumptions made by the SNM are stronger than those made by the MSM. The SNM explicitly models treatment by covariate interaction (and in this example assumes that there is none), whereas the MSM is agnostic about this interaction. G-estimation of the SNM exploits this to gain efficiency, whereas the MSM is more robust since it will be correctly specified regardless of whether or not such an interaction exists.

6.5. G-estimation applied to simulated dataset I: nonparametric equivalence

In our first simulated dataset, A_0, A_1 and L_1 are all binary, and we can specify our SNMM as mentioned previously, except that this is now nonparametric and thus guaranteed to be correctly specified:

$$\psi_1(a_0, a_1, l_1; \phi) = (\phi_0 + \phi_1 a_0 + \phi_2 l_1 + \phi_3 a_0 l_1) a_1 \tag{33}$$

and

$$\psi_0(a_0; \phi) = \phi_4 a_0. \tag{34}$$

In the nonparametric setting, we can find the g-estimates of $\boldsymbol{\phi}$ by direct arithmetic calculation, as was shown by Robins and Hernán [4]. First, we calculate the average outcome Y in each of the eight strata defined by A_0, L_1 and A_1. Using these stratum-specific means and the SNM defined by (33) and (34), we calculate the expected value of the treatment-free counterfactual $Y_0^*(\boldsymbol{\phi})$ in each stratum (as a function of $\boldsymbol{\phi}$). Table IX shows these. The 'no unmeasured confounding' assumption states that, at the true values of $\boldsymbol{\phi}$, $Y_0^*(\boldsymbol{\phi})$ is independent of A_1 within the four strata defined by A_0 and L_1. This means that the average value of $Y_0^*(\boldsymbol{\phi})$ should be the same in lines 1 and 2 of Table IX. That is,

$$2.0114 = 1.2802 - \hat{\phi}_0 \Rightarrow \hat{\phi}_0 = -0.7312.$$

Likewise, the average value of Y_0^* should be the same in lines 3 and 4, the same in lines 5 and 6 and the same in lines 7 and 8. That is,

$$2.3821 = 1.6069 - \hat{\phi}_0 - \hat{\phi}_2 \Rightarrow \hat{\phi}_2 = -0.0440,$$
$$1.4161 - \hat{\phi}_4 = 1.969405 - \hat{\phi}_0 - \hat{\phi}_1 - \hat{\phi}_4 \Rightarrow \hat{\phi}_1 = 0.2904, \text{ and}$$
$$2.063790 - \hat{\phi}_4 = 0.9753 - \hat{\phi}_0 - \hat{\phi}_1 - \hat{\phi}_2 - \hat{\phi}_3 - \hat{\phi}_4 \Rightarrow \hat{\phi}_3 = -0.1244.$$

Table IX. The treatment-free counterfactuals for the g-estimation analysis of the nonparametric structural nested mean model for simulated dataset I.

n	A_0	L_1	A_1	$E(Y)$	$E\left(Y_0^*\right)$
505	0	0	0	2.0011	2.0011
337	0	0	1	1.2802	$1.2802-\phi_0$
159	0	1	0	2.3821	2.3821
15	0	1	1	1.6069	$1.6069-\phi_0-\phi_2$
55	1	0	0	1.4161	$1.4161-\phi_4$
463	1	0	1	0.9753	$0.9753-\phi_0-\phi_1-\phi_4$
385	1	1	0	1.7522	$1.7522-\phi_4$
81	1	1	1	1.1430	$1.1430-\phi_0-\phi_1-\phi_2-\phi_3-\phi_4$

Finally, to estimate ϕ_4, we use the fact that $Y_0^*(\boldsymbol{\phi})$ is independent of A_0. This implies that the average value of $Y_0^*(\boldsymbol{\phi})$ is the same in both the top and bottom halves of Table IX. The average of $Y_0^*(\boldsymbol{\phi})$ in lines 1–4 of the table is given by

$$\frac{505}{1016}\times 2.0114+\frac{337}{1016}\times(1.2802-\phi_0)+\frac{159}{1016}\times 2.3821+\frac{15}{1016}\times(1.6069-\phi_0-\phi_2). \tag{35}$$

The average of Y_0^* in lines 5–8 of the table is given by

$$\frac{55}{984}\times(1.4161-\phi_4)+\frac{463}{984}\times(0.9753-\phi_0-\phi_1-\phi_4)+\frac{385}{984}\times(1.7522-\phi_4)$$
$$+\frac{81}{984}\times(1.1430-\phi_0-\phi_1-\phi_2-\phi_3-\phi_4). \tag{36}$$

Substituting the aforementioned estimates $\hat{\phi}_0$–$\hat{\phi}_3$ for ϕ_0–ϕ_3 and equating (35) and (36) gives

$$\hat{\phi}_4=-0.4996.$$

To obtain estimates of the standard errors of $\hat{\boldsymbol{\phi}}$, we use the bootstrap. The code for this analysis appears in Section A9 of the Supporting Information and the results appear in Table X.

Ideally, to confirm that all three methods give the same estimates in the nonparametric setting, we would like to be able to compare the results of Table X with those of Tables III and V, that is, we would like to use g-estimation to obtain estimates of the MSM in Equation (4) rather than just the parameters of the SNMM defined by the blip functions (33) and (34). Because we have included a treatment-by-confounder interaction in our SNMM, this is not immediately possible for all parameters.

We can certainly estimate $E\left\{Y^{(0,0)}\right\}$: this is the average of the treatment-free counterfactuals, that is, the average of the right-hand column of Table IX. By (34) and (26), we can also estimate $E\left\{Y^{(1,0)}\right\}$ as $E\left\{Y^{(0,0)}\right\}+\hat{\phi}_4$. But by (33) and (26), in order to estimate $E\left\{Y^{(0,1)}\right\}$, we must first obtain an estimate of $E(L_1|A_0=0)$. Then, we have

$$\hat{E}\left\{Y^{(0,1)}\right\}=\hat{E}\left\{Y^{(0,0)}\right\}+\hat{\phi}_0+\hat{\phi}_2\hat{E}(L_1|A_0=0).$$

Table X. The results of the analysis of simulated dataset I, as analysed using g-estimation of the nonparametric structural nested mean model defined by the blip functions (33) and (34), with bootstrap standard errors; the 95% CIs are based on a normal approximation, using the bootstrap standard errors.

Parameter	G-estimation estimate	Bootstrap SE	95% CI	
ϕ_0	−0.7312	0.0373	−0.8043	−0.6580
ϕ_1	0.2904	0.0865	0.1209	0.4598
ϕ_2	−0.0440	0.1197	−0.2786	0.1907
ϕ_3	−0.1244	0.1545	−0.4273	0.1784
ϕ_4	−0.4996	0.0473	−0.5923	−0.4069

Likewise, we can estimate $E\left\{Y^{(0,1)}\right\}$ as

$$\hat{E}\left\{Y^{(1,1)}\right\} = \hat{E}\left\{Y^{(0,0)}\right\} + \hat{\phi}_0 + \hat{\phi}_1 + \left(\hat{\phi}_2 + \hat{\phi}_3\right)\hat{E}(L_1|A_0 = 1).$$

For $\hat{E}(L_1|A_0 = 0)$ we substitute the proportion of subjects with $L_1 = 1$ out of those for whom $A_0 = 0$, that is, 174/1016. And similarly for $\hat{E}(L_1|A_0 = 1)$, we substitute the proportion of subjects with $L_1 = 1$ out of those for whom $A_0 = 1$, that is, 466/984. This leads to the g-estimates of each of the four potential outcomes and hence (by (5)) the g-estimates of the parameters of the MSM (4). Again, we use the bootstrap to obtain standard errors. The code appears in Section A9 of the Supporting Information, and the results are numerically identical to those shown in Table III.

6.6. Loss to follow-up

Under the assumption that missingness is at random (MAR), that is, that

$$P\left(R_{t,i} = 1 \,\middle|\, \bar{A}_i, \bar{L}_i, R_{t-1,i} = 1\right) = P\left(R_{t,i} = 1 \,\middle|\, \bar{A}_{t-1,i}, \bar{L}_{t-1,i}, R_{t-1,i} = 1\right),$$

loss to follow-up can be dealt with in g-estimation of SNMs by re-weighting the contributions made by each incompletely observed subject. The following weights are defined:

$$\Omega_{t,i} = \frac{1}{\prod_{s=0}^{t} P\left(R_{s,i} = 1 \,\middle|\, \bar{A}_{s-1,i}, \bar{L}_{s-1,i}, R_{s-1,i} = 1\right)}.$$

We carry out the g-estimation procedure on the re-weighted available data as follows. For each t, we fit the model for A_t conditional on past treatment and covariate histories only to those with A_t observed, re-weighting the subjects' contributions by $\Omega_{t,i}$. In the simplest approach, we include only those with full data in the estimating equation, and these are re-weighted by $\Omega_{T,i}$, although a more efficient approach would augment this estimating equation to include contributions by those lost to follow-up [32]. We provide a worked example in Section H of the Supporting Information.

7. Some extensions

We have seen that, in the special case where both time-changing treatment and time-changing confounder are binary (or more generally, categorical) and the number of visits is low, it is possible to apply all three approaches (g-computation formula, IPW and g-estimation) nonparametrically, and the estimates are identical. However, as we move to more realistic settings with non-saturated models, the parametric models that need to be specified for each method are different. For the g-computation formula, we specify a model for Y conditional on $\bar{L}$ and $\bar{A}$ and a model for each L_t in $\bar{L}$, conditional on past history $\bar{L}_{t-1}$ and $\bar{A}_{t-1}$. For IPW, we specify models for each A_t in $\bar{A}$, conditional on past history $\bar{L}_t$ and $\bar{A}_{t-1}$. Doubly robust (DR) estimators [33] combine both these approaches to give estimators of the parameters of an MSM, which remain consistent if at least one of the sets of models (either those needed for g-computation or those needed for IPW, or both) has been correctly specified, thus offering some protection against model misspecification. The standard DR estimators are also guaranteed to be asymptotically more efficient than IPW estimators whenever the models for Y and L_t are correctly specified. Intuitively, one can think of DR estimators as offering the 'best of both worlds', combining the efficiency of the g-computation formula with the robustness of IPW. However, when the inverse probability weights are too variable, the DR estimator can still inherit much of the associated imprecision and instability, and when the models for Y and L_t are misspecified, DR estimators can be less efficient than IPW estimators, even when this misspecification is quite mild [26, 34]. In response to these criticisms, improved DR estimators have been developed that address these problems [35–38]. DR estimators have also been proposed for estimating the parameters of an SNM [34].

We have discussed in this tutorial methods for comparing static treatment regimes, such as $\bar{a} = (0, 0)$ and $\bar{a} = (1, 0)$. Although in the observational data at hand, the subjects' treatment histories depend on their covariate histories, our focus has always been on getting rid of this dependence. As was discussed in Section 2.7, in many situations, the practical question of interest asks for a comparison of dynamic regimes, such as 'treat until patient becomes anaemic, then stop treatment'. All the methods discussed in this tutorial have been extended to compare dynamic regimes [4, 39–43].

8. Summary: strengths and limitations of each of the methods

8.1. G-computation formula

8.1.1. Strengths.

1. This approach can be conceptualised as the extension of standardisation to treatments applied over many timepoints. The missing counterfactuals are predicted, leading naturally to a comparison of 'what would have happened' to the whole population under different hypothetical interventions. This is also possible with IPW of MSMs but is not so immediate for g-estimation of SNMs.
2. The method can be implemented using routines in SAS and STATA.
3. As has been shown by Taubman *et al.* [19], the method can easily handle complex joint interventions (such as—to take an example from this reference—'do not smoke *and* exercise at least 30 minutes a day') and is particularly suited to situations in which a relatively small number of interventions are to be compared.
4. The method can also cope well with different sorts (continuous, binary, categorical, time-to-event, etc.) of outcome variable.
5. If applied nonparametrically, no additional assumptions besides the 'no unmeasured confounding' assumption are needed. More typically, however, parametric models must be specified for the outcome conditional on the covariate and treatment histories and for the time-changing covariate(s) at a given visit conditional on the covariate and treatment histories up to that visit. This almost fully parametric specification (everything except for the treatment is modelled), as long as all models are approximately correctly specified, leads to increased statistical efficiency. Thus, if the choice of these models can be justified from sound prior knowledge of the relationships between the variables, the method can perform very well.
6. Each hypothetical intervention can be compared with a baseline of 'no intervention', where by no intervention, we mean that every subject's treatment value would be as it was in the observational setting, that is, $(a_0, a_1, \ldots, a_T) = (A_0, A_1 \ldots, A_T)$. This is often a very useful comparison for informing public health policy, because the likely impact of the considered interventions can be compared with maintaining the status quo [44]. This can be carried out by simply comparing the mean potential outcome under the hypothetical intervention with the mean of the actual observed outcomes, and the same comparison could be made using IPW of MSMs. Alternatively, the comparison can be made by additionally specifying models for each A_t conditional on $\bar{A}_{t-1}$ and $\bar{L}_t$. The approach is then fully parametric in the sense that a model for the joint distribution of all post-baseline variables is specified. If this joint distribution is correctly specified and the 'no unmeasured confounding' assumption holds, then the mean of the potential outcomes under 'no intervention' should be equal (except for finite-sample error) to the mean of the actual observed outcomes. Thus, a comparison of the two acts as a check (although not definitive) of the validity of the assumptions. See [19] for further details.

8.1.2. Limitations.

1. When considering continuous treatments, categorical treatments with many levels or even binary treatments with many timepoints, the fact that the g-computation algorithm is really a method for estimating $E\left(Y^{\bar{a}}\right)$ for each possible $\bar{a}$ can make interpretation difficult, because there are simply too many possible hypothetical interventions to be compared. If there are a few hypothetical interventions of interest *a priori*, then the potential outcomes associated with these can be compared. Otherwise, a parametric model also needs to be specified for $E\left(Y^{\bar{a}}\right)$ in terms of $\bar{a}$ (i.e. an MSM) so that the comparison can be summarised in terms of fewer parameters. In this case, the fact that the estimation of the parameters of the MSM is *post hoc* is a drawback compared with the other two methods, where the estimation of the parameters of the structural model of interest is directly targeted. Furthermore, there is the issue that this model may be incompatible with the specified parametric associational models, as discussed in Section 5.1.3.
2. The g-computation formula is computationally intensive. Monte Carlo simulation and bootstrapping are both required in any realistic setting.
3. In the parametric setting, when the time-varying covariate is binary and when the models for L_t given $\left(\bar{L}_{t-1}, \bar{A}_{t-1}\right)$ are not saturated, some common modelling choices give rise to the

g-null paradox: that the model is inconsistent with the causal null hypothesis whenever there is a non-null effect of past treatment on future covariates.
4. Even if the causal null hypothesis can be ruled out *a priori*, meaning that the g-null paradox is not an issue, the (near) fully parametric specification needed for the g-computation formula means that model misspecification remains a serious concern, particularly when L_t is high-dimensional and g-computation (unlike the other two approaches) requires correctly specifying its joint distribution given $\left(\bar{L}_{t-1}, \bar{A}_{t-1}\right)$. In the parametric version, great care should always be taken to ensure that our estimates of $E\left(Y^{\bar{a}}\right)$ for some values of $\bar{a}$ do not rely on extrapolation beyond the range of the observed data.

8.2. *Inverse probability weighting estimation of marginal structural models*

8.2.1. *Strengths.*

1. Re-weighting to achieve a pseudo-sample unaffected by confounding is intuitive and relatively easy to explain. Of the three methods, it is the most closely related to standard methods.
2. The analysis is easily implementable using weighted versions of standard routines.
3. As with the g-computation formula but in contrast with g-estimation of SNMs, MSMs cope well with different sorts of outcome variable. It is possible to fit a logistic MSM for binary outcomes and a Cox MSM for time-to-event outcomes.
4. Only models for the treatment assignment and the MSM itself need be specified. The conditional distribution of Y given the covariates and the conditional distribution of the covariates given past covariates and treatments are left unspecified. This makes the method less prone to model misspecification than the g-computation formula, and the difference becomes increasingly important as the dimension of the covariates increases.
5. The g-null paradox does not arise.
6. There is a clear set of parameters $\boldsymbol{\gamma}$ (excluding γ_0) such that if all these parameters are zero, then the causal null hypothesis holds, and these parameters are the direct target of estimation.
7. Possible interactions between treatment and time-varying covariates remain unspecified, and thus, provided that these are not of interest, it is an advantage that there is no possibility of incorrectly forgetting or misspecifying them.

8.2.2. *Limitations.*

1. Inverse weighting can be unstable and inefficient if there are extreme weights. This problem can be partially mitigated by using stabilised weights.
2. Continuous treatments are difficult to handle.
3. IPW cannot be used when the experimental treatment assignment assumption is violated. For example, in our analysis of simulated dataset I, had all subjects for whom $A_0 = L_1 = 0$ remained untreated at visit 1, we would have no information on $E(Y|A_0 = 0, L_1 = 0, A_1 = 1)$, and no amount of re-weighting could recover this information, and thus—for example—$E\left\{Y^{(0,1)}\right\}$ would not be estimable. The other methods, by proposing parametric models for $E(Y|A_0, L_1, A_1)$, can recover this information but only by using these models to extrapolate beyond the data.
4. Possible interactions between treatment and time-varying covariates cannot be explored because the MSM is marginal with respect to the latter.

8.3. *G-estimation of structural nested models*

8.3.1. *Strengths.*

1. G-estimation is usually more efficient than IPW [34] but requires fewer parametric assumptions than the g-computation formula. In this sense, it may be seen as a good compromise with regard to the bias–variance trade-off.
2. Only models for the treatment assignment and the 'blipping down' process need be specified.
3. Interactions between treatment and time-varying covariates can be incorporated into the model.
4. The g-null paradox does not arise.
5. There is a clear set of parameters $\boldsymbol{\phi}$ such that $\boldsymbol{\phi} = \mathbf{0}$ corresponds to the causal null hypothesis, and these parameters are the direct target of estimation.

6. The conditional nature of the causal effects defined via an SNM mean that they are more easily transportable to different populations [45].

8.3.2. Limitations.

1. G-estimation is not currently implementable using standard software in all settings. Although stgest in STATA (for structural nested failure time models) and SURV in SAS are both useful macros, they do not cover the whole range of SNMs.
2. For binary data, there is no SNMM with estimable parameters that respects the fact that probabilities must lie between 0 and 1. However, a new set of models suggested by Vansteelandt [46] promises to resolve this issue.
3. For survival data, an extension of the accelerated failure time model must be used. This is not the most commonly used of survival models and thus might deter a potential user of these methods.
4. For survival data with administrative censoring, when a structural nested accelerated failure time model may be used, a closed-form solution such as given in (30) cannot be obtained, and thus a searching algorithm must be used. This is computationally intensive, and thus to make the estimation feasible, strong restrictions on the parameters of the blipping down model are usually needed. This increases the possibility of misspecifying the SNM itself. Note however that an alternative approach based on structural nested cumulative failure time models has recently been advocated, which circumvents this problem [47].

9. Concluding remarks

In this tutorial, we first highlighted the potential problems associated with using standard regression methods to control for time-varying confounders. Robins and others have proposed three methods that avoid these problems, provided that some structural and parametric modelling assumptions hold. We have described the application of these methods to simple simulated examples.

To highlight the roles played by various modelling assumptions, we first considered an example in which all methods could be applied nonparametrically and confirmed that the results from all three methods were identical. We then considered a setting in which the different modelling assumptions made by the three methods led to differences in the estimates and in the statistical efficiency of the estimators. We have discussed the strengths and limitations of these methods in terms of statistical efficiency, robustness to modelling assumptions, numerical stability and extensions to more complicated settings and to various types of outcome data.

As well as dealing with time-dependent confounding, the methods can all be adapted to deal with loss to follow-up, under the MAR assumption, and we have provided further examples in the Supporting Information to illustrate this.

Acknowledgements

This work was supported by grant G0701024 from the Medical Research Council, U.K. The authors wish to thank three anonymous referees for their enlightening and careful comments, which have greatly improved the tutorial.

References

1. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; **7**:1393–1512.
2. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
3. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; **3**:319–336.
4. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (eds). Chapman and Hall/CRC Press: New York, 2009; 553–599.
5. Robins JM. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 1992; **79**(2):321–334.
6. Witteman JCM, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MAJ, Hofman A, Robins JM. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *American Journal of Epidemiology* 1998; **148**(4):390–401.

7. Tilling K, Sterne JA, Szklo M. Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-estimation: the atherosclerosis risk in communities study. *American Journal of Epidemiology* 2002; **155**(8):710–718.
8. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**:561–570.
9. Rubin DB. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* 1978; **6**:34–58.
10. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; **20**(6):880–883.
11. Ten Have TR, Joffe MM. A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research* 2012; **21**(1):77–107.
12. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiological research. *Epidemiology* 1999; **10**(1):37–48.
13. Pearl J. Causal inference in the Health Sciences: a conceptual introduction. *Health Services & Outcomes Research Methodology* 2001; **2**:189–220.
14. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–625.
15. Robins JM, Wasserman L. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Providence Rhode Island*, Geiger D, Shenoy P (eds). Morgan Kaufmann: San Francisco, 1997; 409–420.
16. Shvidel L, Sigler E. Symptomatic anemia induced by rosiglitazone. *European Journal of Internal Medicine* 2007; **18**(4):348.
17. Daniel RM, De Stavola BL, Cousens SN. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal* 2011; **11**(4):479–517.
18. Neugebauer R, van der Laan MJ. Causal effects in longitudinal studies: definition and maximum likelihood estimation. *Computational Statistics & Data Analysis* 2006; **51**(3):1664–1675.
19. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology* 2009; **38**(6):1599–1611.
20. Rubin DB. Inference and missing data (with discussion). *Biometrika* 1976; **63**:581–592.
21. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**:656–664.
22. Fewell Z, Hernán MA, Wolfe F, Tilling K, Choi H, Sterne JAC. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal* 2004; **4**:402–420.
23. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* 2011; **173**(7):731–738.
24. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**(1):29–46.
25. Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B* 1951; **13**:238–241.
26. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science* 2007; **22**(4):523–580.
27. Tan Z. Understanding OR, PS and DR. Comment on "Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data" by Kang and Schafer. *Statistical Science* 2007; **22**(4):560–568.
28. Robins JM. Analytic methods for estimating HIV-treatment and cofactor effects. In *Methodological Issues in AIDS Behavioral Research*, Ostrow DG, Kessler RC (eds). Plenum Press: New York, 1993; 213–290.
29. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics—Theory and Methods* 1994; **23**:2379–2412.
30. Casella G, Berger RL. *Statistical Inference*, 2nd ed. Wadsworth & Brooks: Pacific Grove, CA, 2002.
31. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal* 2002; **2**:164–182.
32. Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer: New York, 2006.
33. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**(4):962–973.
34. Goetgeluk S, Vansteelandt S, Goetghebeur E. Estimation of controlled direct effects. *Journal of the Royal Statistical Society Series B* 2008; **70**:1049–1066.
35. Tan Z. Comment: Improved local efficiency and double robustness. *The International Journal of Biostatistics* 2008; **4**. Article 10.
36. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 2010; **97**:661–682.
37. Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 2009; **96**:723–734.
38. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* 2012; **21**(1):7–30.
39. Robins JM, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* 2008; **27**:4678–4721.
40. Robins JM. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, Lin DY, Heagerty P (eds). Springer: New York, 2004; 189–326.
41. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology* 2006; **98**(3):237–242.
42. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: main content. *The International Journal of Biostatistics* 2010; **6**(2). Article 8.

43. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics* 2010; **6**(2). Article 18.
44. Hubbard AE, van der Laan MJ. Population intervention models in causal inference. *Biometrika* 2008; **95**(1):35–47.
45. Vansteelandt S, Keiding N. Invited commentary: G-computation—lost in translation? *American Journal of Epidemiology* 2011; **173**(7):739–742.
46. Vansteelandt S. Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models. *Biometrika* 2010; **97**(4):921–934.
47. Picciotto S, Hernán MA, Page JH, Young JG, Robins JM. Structural nested cumulative failure time models to estimate the effects of interventions. *Journal of the American Statistical Association* 2012; **107**(499):886–900.