Causal Inference for Complex Longitudinal Data: the continuous case.

Richard D. Gill and James M. Robins
Mathematical Institute, University Utrecht and Eurandom, Eindhoven;
Depts. of Epidemiology and Biostatistics, Harvard School of Public Health
14 March 1999

Abstract.
We extend Robins' theory of causal inference for complex longitudinal data to the case of continuously varying as opposed to discrete covariates and treatments. In particular we establish versions of the key results of the discrete theory: the g-computation formula and a collection of powerful characterizations of the g-null hypothesis of no treatment effect. This is accomplished under natural continuity hypotheses concerning the conditional distributions of the outcome variable and of the covariates given the past.

1. The problem.

Robins (1986, 1987, 1989, 1997) introduced the following framework for describing a longitudinal observational study in which new treatment decisions are repeatedly taken on the basis of accumulating data. Suppose a patient will visit a clinic at K time points. At visit $k = 1, \ldots, K$, medical tests are done yielding some data L_k. The data $L_1, \ldots, L_{k-1}$ from earlier visits is still available. The doctor gives a treatment A_k (this could be the quantity of a certain drug). Earlier treatments $A_1,\ldots,A_{k-1}$ are also known. Of interest is some response Y, to be thought of as representing the state of the patient after the complete treatment. Thus in time sequence the complete history of the patient results in the alternating sequence of covariates (or responses) and treatments

$$L_1, A_1, \ldots, L_K, A_K, Y.$$

Any of the variables may be vectors and may take values in different spaces. The notation L_k for covariate and A_k for treatment was inspired by AIDS studies where L_k is lymphocyte count (white blood corpuscles) and A_k is the dose of the drug AZT at the k'th visit to the clinic. Robins' approach generalizes the time-independent point-treatment counterfactual approach of Neyman (1923) and Rubin (1974, 1978, 1983) to the setting of longitudinal studies with time-varying treatments and covariates. Robins (1995, 1997) discusses the relationship between his theory and causal theories based on directed acyclic graphs and non-parametric structural equation models due to Pearl (1995) and Spirtes, Glymour, and Scheines (1993).

The study typically yields values of an i.i.d. sample of this collection of random variables. On the basis of this data we want to decide whether treatment influences the final outcome Y, and if so, how. In this paper we do not however consider statistical issues, but concentrate on identification and modelling questions. We take the joint probability distribution of the data $(L_1, A_1, \ldots, L_K, A_K, Y)$ as being given and ask whether the effect of treatment is identified, when this distribution is known.

Note that we are considering an observational study, not a planned clinical trial. The treatment decision at the k'th visit is not determined by a specified protocol but is the result of the doctor's personal decision at that moment. In different instances the treatment A_k given at the kth visit will vary even though the available information $L_1, A_1, \ldots, A_{k-1}, L_k$ is the same. Indeed, it is precisely this variation which will allow us to study the effect of treatment on outcome.

In Robins' theory (some parts of which are presented below) the covariates and treatments take values in discrete spaces. Our aim here is to extend the theory to the general case. One might argue that in practice all data is discrete, but still in practice one will often want to work with continuous models. Our original motivation was to rigorously develop Robins' (1997) outline of a theory of causal inference when treatments and covariates can be administered and observed continuously in time. Here again it is necessary to face up to the same questions, if the theory is to be given a firm mathematical foundation.

Write $\overline{L}_k = (L_1, \ldots, L_k)$, $\overline{A}_k = (A_1, \ldots, A_k)$; we abbreviate $\overline{L}_K$ and $\overline{A}_K$ to $\overline{L}$ and $\overline{A}$. Values of the random variables are denoted by the corresponding lower case letters. The aim is to decide how a specified treatment regime would affect outcome. A treatment regime or plan, denoted g, is a rule which specifies treatment at each time point, given the data available at that moment. In other words it is a collection (g_k) of functions g_k, the k'th defined on sequences of the first k covariate values, where $a_k = g_k(\overline{l}_k)$ is the treatment to be administered at the k'th visit given covariate values $\overline{l}_k = (l_1, \ldots, l_k)$ up till then. Following the notational conventions already introduced, we define $\overline{g}_k(\overline{l}_k) = (g_1(l_1), g_2(l_1, l_2), \ldots, g_k(l_1, \ldots, l_k))$ and $\overline{g}(\overline{l}) = \overline{g}_K(\overline{l}_K)$. However for brevity we often abbreviate $\overline{g}_k$ or $\overline{g}$ simply to g when the context makes clear which function is meant, as in $\overline{a}_k = g(\overline{l}_k)$ or $\overline{a} = g(\overline{l})$.

Robins' approach is to assume that for given g is defined, alongside of the 'factual' $(\overline{L}, \overline{A}, Y)$, another so-called counterfactual random variable Y^g: the outcome which would have been obtained if the patient had actually been treated according to the regime g. His strategy is to show that the probability distribution of the counterfactual Y^g can be recovered from that of the factual $(\overline{L}, \overline{A}, Y)$ under some assumptions on the joint distribution of $(\overline{L}, \overline{A}, Y)$ and Y^g. Assuming all variables are discrete, his assumptions are:

A1: Consistency. $Y = Y^g$ on $\overline{A} = g(\overline{L})$.

A2: Randomization. $A_k \perp Y^g \mid \overline{L}_k, \overline{A}_{k-1}$ on $\overline{A}_{k-1} = g(\overline{L}_{k-1})$.

A3: Evaluability. For each k and $\overline{a}_k$, $\overline{l}_k$ with $\overline{a}_k = g(\overline{l}_k)$, $\Pr(\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) > 0 \Rightarrow \Pr(\overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k) > 0$.

The consistency assumption A1 states that if a patient coincidentally is given the same sequence of treatments as the plan g would have prescribed, then the outcome is the same as it would have been under the plan. The randomisation assumption A2 states that the k'th assignment of treatment, given the information available at that moment, does not depend on the future outcome under the hypothetical plan g. This assumption would be true if treatment was actually assigned by randomization as in a controlled sequential trial.

On the other hand it would typically not be true if the doctor's treatment decisions were based on further variables than those actually measured which gave strong indications of the patient's underlying health status (and hence likely outcome under different treatment plans). The evaluability condition A3 states that the plan g was in a sense actually tested in the factual experiment: when there was an opportunity to apply the plan, that opportunity was at least sometimes taken.

Under these conditions the distribution of Y^g can be computed by the g-computation formula:

$$\Pr(Y^g \in \cdot) = \int\limits_{l_1; a_1 = g_1(l_1)} \cdots \int\limits_{l_K; a_K = g_K(\overline{l}_K)} \Pr(Y \in \cdot \mid \overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K) \prod_{k=1}^{K} \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}). \tag{1}$$

Moreover, the right-hand side is a functional of the joint distribution of the factual variables only and of the chosen treatment plan g, and we sometimes refer to it as $b(g)$ or $b(g; \mathrm{law}(\overline{L}, \overline{A}, Y))$. In particular, it does not involve conditional probabilities for which the conditioning event has zero probability. We indicate the proof in a moment; it is rather straightfoward formula manipulation. First we discuss some interpretational issues.

In practice computation of the right hand side of (1) could be implemented by a Monte-Carlo experiment, as follows. An asterix is used to denote the simulated variables. First set $L_1^* = l_1^*$ drawn from the marginal distribution of L_1. Then set $A_1^* = a_1^* = g_1(l_1^*)$. Next set $L_2^* = l_2^*$ drawn from the conditional distribution of L_2 given $L_1 = l_1^*, A_1 = a_1^*$; and so on. Finally set $Y^* = y^*$ drawn from the conditional distribution of Y given $\overline{L} = \overline{l}^*, \overline{A} = \overline{a}^*$.

This *probabilistic* reading of (1) begs a subject-matter interpretation in terms of further counterfactual variables: the outcomes L_k^g of the k'th covariate, when patients are treated by plan g. It seems as if we believe that

B1: the distribution of L_k^g given the (counterfactual) past, is the same as that of L_k given the same values of the factual variables.

However this interpretation is only valid under additional assumptions. Specifically, if we can add to A2

A2†: Causal graph. $A_k \perp (Y^g, L_{k+1}^g, \ldots, L_K^g) \mid \overline{L}_k, \overline{A}_{k-1}$ on $\overline{A}_{k-1} = g(\overline{L}_{k-1})$

then one can *prove* it by an argument on the same lines as that which proves (1).

It is important to note that we do not need assumption A2† in proving (1), and hence that (1) can be valid without its obvious probabilistic interpretation B1 being correct. Note A2† would hold in a sequential randomized trial. However, in an observational study, A2 may be true but A2† false. For example, Robins (1997, pp. 81–83) describes a substantively plausible data-generating mechanism which depends on further unobserved variables U_m, and under which, for certain choices of g, assumption A2 is true but assumption A2† is false, once the U_m have been integrated out. We are convinced by such examples that (1) should not be regarded as the definition of $\Pr(Y^g \in \cdot)$ but rather needs to be derived from the more primitive conditions A1 to A3. We believe that these conditions are both meaningful and as weak as possible. Hence our programme to generalise to continuous variables is also important.

The proof of (1) is as follows. Consider the right hand side of (1). By assumption A1 we may replace Y by Y^g in the conditional probability which is the integrand of this expression. Now repeatedly carry out the following operations: using A2 drop the last conditioning variable "$A_K = a_K$" from the integrand. Next integrate out over l_K, so that the K'th term in the product of conditional distributions disappears and the conditioning on $L_K = l_K$ in the integrand is also dropped. Now the right hand side of (1) (but with Y^g in place of Y) has been transformed into the same expression with K replaced by $K-1$. Repeat these steps of dropping the last a_k and integrating out the last l_k another $K-1$ times and finally the left hand side of (1) is obtained.

Note that this proof of (1) only uses assumptions A1 and A2. Assumption A3 can be used (in a similarly easy argument) to show that the right-hand side of (1) is uniquely defined, i.e., independently of choice of conditional probabilities given zero probability events. But where are the problems in going to the continuous case? Our proof of (1) using A1 and A2 seemed to be perfectly general.

The problem is that when the treatments $\overline{A}$ are continuously distributed, the set of $(\overline{l}_k, \overline{a}_k)$ which are of the form $(\overline{l}_k, \overline{g}_k(\overline{l}_k))$ for a particular g will be a zero probability set for $(\overline{L}_k, \overline{A}_k)$. Hence the events referred to in A1 and A2 are zero probability events in the continuous case, and the conditional distributions on the right-hand side of (1) are only needed on these zero probability events. They can be chosen arbitrarily, making the right-hand side of (1) more or less arbitrary. Perhaps they can be chosen in order to make (1) correct, but then we need to know how to pick the right versions. Thus A1 and A2 need to be strengthened somehow for a meaningful theory. As it stands, Condition A3 is empty in the continuous case, but a reformulation of it in terms of supports of the distributions involved will turn out to do the same job.

In this paper we will make some natural continuity assumptions which give us a preferred choice of conditional distributions. Then we answer the questions: is equation (1) *correct*, and is the right-hand side *uniquely* determined by the joint distribution of the factuals? The three assumptions A1 to A3 will be reformulated to take account of the new context, and the proof of (1) will no longer be a completely trivial exercise though it still follows the same line as given above.

We go on to investigate whether the key theorems in Robins' (1986, 1987, 1989, 1997) theory of causal inference for complex longitudinal data remain valid in the new context.

A further type of question we want to consider is the following: given factual variables $(\overline{L}, \overline{A}, Y)$ can one construct a variable Y^g satisfying A1–A2? If this were not the case, then the assumption of existence of the counterfactuals places restrictions on the distribution of the data. If on the other had it is true, then the often heated discussion about whether or not counterfactual reasoning makes sense loses a major part of its sting: as a thought experiment we can always suppose the counterfactuals exist. If this leads us to useful statistical models and analysis techniques, that is fine.

We emphasize that the correctness of (1), and the uniqueness of (the right-hand side) of (1), are two different issues. It is possible to construct simple examples where there are two different counterfactual variables Y^g and Y'^g, with different marginal distributions, both satisfying A1–A2, but with different versions of conditional distributions; in each case the right-hand side of (1) gives the 'right' answer if the 'right' choice of conditional distributions is taken. Here is such an example with $K = 1$; L_1 trivial; so there are only two factual variables $A = A_1$ and Y under consideration. Let the sample space Ω be the unit interval with the uniform probability distribution on it, but with an extra point (of zero probability) $\frac{1}{2}'$ immediately after the point $\frac{1}{2}$. Let $A(\omega) \equiv \omega$ (with $A(\frac{1}{2}) = A(\frac{1}{2}') = \frac{1}{2}$), and let $Y(\omega) = 0$ for $\omega \leq \frac{1}{2}$, $Y(\omega) = 1$ for $\omega \geq \frac{1}{2}'$. Let the treatment g be the fixed value $a = \frac{1}{2}$, and let $Y^g \equiv 1$ except that $Y^g(\frac{1}{2}) = 0$; let $Y'^g \equiv 0$ except that $Y'^g(\frac{1}{2}') = 1$. Note that $Y^g = Y'^g = Y$ on $\frac{1}{2}, \frac{1}{2}'\} = \{A = \frac{1}{2}\}$. Furthemore, Y^g and Y'^g both have degenerate distributions so are trivially independent of A. Thus conditions A1 and A2 hold for both Y^g and $Y'g$. Choosing the conditional distribution of Y given $A = \frac{1}{2}$ either to be degenerate at 1 or degenerate at 0 produces the 'right' answer for each of the two counterfactuals. What is going on here is that the distribution of the data cannot possibly tell us what the result of the treatment $a = \frac{1}{2}$ should be. We have two equally plausible counterfactuals Y^g and Y'^g satisfying all our conditions but with completely different distributions. The law of Y given $A = \frac{1}{2}$ could reasonably be taken to be almost anything. However the law of Y given other values of A seems more well-defined. In fact it can be chosen to be continuous in a (except at $a = \frac{1}{2}$ and the choice subject to continuity seems compelling.

Our approach will be to assume that the conditional distributions involved can be chosen in a continous way—continuous, in the sense of weak convergence, as the values of the conditioning variables vary throughout their support. It then turns out that if one chooses versions of conditional distributions subject to continuity, there is in fact no choice: the continuous version is uniquely defined. Formula (1) will now be uniquely defined, under a natural restatement of A3, and when choosing the conditional distributions appearing in the formula subject to continuity. The question whether or not it gives the right answer requires parallel continuity assumptions concerning the distribution of the counterfactual outcome given factual variables.

At the end of the paper we will pay some attention to an alternative approach. We replace the idea of a treatment plan assigning a fixed amount of treatment given the past, by a plan where the amount of treatment given the past stays random. This seems very natural since even if a treatment plan nominally calls for a certain exact quantity of some drug to be administered, in practice the amount administered will not be precisely constant. The uniqueness question is very easily solved under a natural restatement of A3. However whether or not the answer is the right answer turns out to be a much more delicate issue and we give a positive answer under a rather different kind of regularity condition, not assuming continuity any more but instead making non-distributional assumptions on the underlying probability space. This approach raises some interesting open problems.

2. Facts on conditioning.

Conditional distributions. We assume without further mention from now on that all variables take values in Polish spaces (i.e., complete separable metric spaces). This ensures, among other things, that conditional distributions of one set of variables given values of other sets *exist*, in other words, letting X and Y denote temporarily two groups of these variables, joint distributions can be represented as

$$\Pr(X \in \mathrm{d}x, Y \in \mathrm{d}y) = \Pr(X \in \mathrm{d}x \mid Y = y) \Pr(Y \in \mathrm{d}y). \tag{2}$$

When we talk about a version of the law of X given Y we mean a family of laws $\Pr(X \in \cdot \mid Y = y)$ satisfying (2).

Repeated conditioning. Given versions of the law of X given Y and Z, and of Y given Z, one can construct a version of the law of X given Z as follows:

$$\int \Pr(X \in \cdot \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z) = \Pr(X \in \cdot \mid Z = z).$$

Fact 4 below shows that if the two conditional distributions on the left hand side are chosen subject to a continuity property, then the result on the right hand side maintains this property.

Conditional independence. When we say that $X \perp Y \mid Z$ we mean that there is a version of the joint laws of (X, Y) given $Z = z$ according to which X and Y are independent for every value z. It follows that any version of the law of X given $Z = z$ supplies a version of the law of X given $Y = y, Z = z$. Conversely, if it is impossible to choose versions of $\mathrm{law}(X \mid Y, Z)$ which for each z do not depend on y, then $X \not\perp Y \mid Z$.

Support of a distribution. We define a support point of the law of X as a point x such that $\Pr(X \in B(x, \delta)) > 0$ for all $\delta > 0$, where $B(x, \delta)$ is the open ball around x of radius δ. We define the support of X to be the set of all support points. As one might expect, it does support the distribution of X, i.e., it has probability one (Fact 1 below).

The following four facts will be needed. The first two are well-known but they are given here including proofs for completeness. The reader may like to continue reading in the next section and only come back here for reference.

Fact 1. *The support of X, $\mathrm{Supp}(X)$, is closed and has probability* 1.

Proof. Any point not in the support is the centre of an open ball of probability zero. All points in this ball are also not support points. The complement of the support is therefore open. By separability it can be expressed as a countable union of balls of probability zero, hence it has probability zero. □

It follows that one can also characterise the support of X as the smallest closed set containing X with probability 1.

Fact 2. *Suppose* law $(X \mid Y = y)$ *can be chosen continuous in* $y \in \mathrm{Supp}(Y)$ *(with respect to weak convergence). Then subject to continuity it is uniquely defined there, and equals* $\lim_{\delta \downarrow 0} \mathrm{law}\,(X \mid Y \in B(y, \delta))$.

Proof. Choose versions of law $(X \mid Y = y)$ subject to continuity. Fix a point $y_0 \in \mathrm{Supp}(Y)$ and let f be a bounded continuous function. Then

$$E(f(X) \mid Y \in B(y_0, \delta)) = \int_{B(y_0,\delta)\cap \mathrm{Supp}(Y)} E(f(X) \mid Y = y) \Pr(Y \in \mathrm{d}y \mid Y \in B(y_0, \delta))$$

where $E(f(X) \mid Y = y)$ inside the integral on the right hand side is computed according to the chosen set of conditional laws. By continuity (with respect to weak convergence) of these distributions, it is a continuous and bounded function of y. Since law$(Y \mid Y \in B(y_0, \delta)) \to \delta_{y_0}$ as $\delta \downarrow 0$, the right hand side converges to $E(f(X) \mid Y = y_0)$ as $\delta \downarrow 0$. □

Fact 3. *Suppose* law$(X \mid Y = y)$ *can be chosen continuous in* $y \in \mathrm{Supp}(Y)$. *Then for* $y \in \mathrm{Supp}(Y)$, $\mathrm{Supp}(X \mid Y = y) \times \{y\} \subseteq \mathrm{Supp}(X, Y)$.

Proof. For $y \in \mathrm{Supp}(Y)$ and $x \in \mathrm{Supp}(X \mid Y = y)$ we have for all $\delta > 0$ since $B(y, \delta)$ is open

$$0 < \Pr(X \in B(x, \delta) \mid Y = y) \leq \liminf_{\varepsilon \downarrow 0} \Pr(X \in B(x, \delta) \mid Y \in B(y, \varepsilon)).$$

So for arbitrary δ and then small enough ε, $\Pr(X \in B(x, \delta) \mid Y \in B(y, \varepsilon)) > 0$, but also $\Pr(Y \in B(y, \varepsilon)) > 0$. But

$$\Pr((X, Y) \in B(x, \delta) \times B(y, \delta)) \geq \Pr(Y \in B(y, \varepsilon)) \Pr(X \in B(x, \delta) \mid Y \in B(y, \varepsilon))$$

for all $\varepsilon < \delta$, which is positive for small enough ε. □

One might expect that the union over $y \in \mathrm{Supp}(Y)$ of the sets $\mathrm{Supp}(X \mid Y = y) \times \{y\}$ is precisely equal to $\mathrm{Supp}(X, Y)$ but this is not necessarily the case. The resulting set can be strictly contained in $\mathrm{Supp}(X, Y)$ though it is *a* support of (X, Y) in the sense of having probability one. Its closure equals $\mathrm{Supp}(X, Y)$.

Fact 4. *Suppose* $\Pr(X \in \cdot \mid Y = y, Z = z)$ *is a family of conditional laws of* X *given* Y *and* Z, *jointly continuous in* $(y, z) \in \mathrm{Supp}(Y, Z)$. *Suppose* $\Pr(Y \in \cdot \mid Z = z)$ *is continuous in* $z \in \mathrm{Supp}(Z)$. *Then*

$$\Pr(X \in \cdot \mid Z = z) = \int_y \Pr(X \in \cdot \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z)$$

is continuous in z.

Proof. Let f be a bounded continuous function, let z_0 be fixed and in the support of Z. We want to show that

$$\int E(f(X) \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z) \\ \to \int E(f(X) \mid Y = y, Z = z_0) \Pr(Y \in \mathrm{d}y \mid Z = z_0)$$

as $z \to z_0$, $z \in \mathrm{Supp}(Z)$. Suppose without loss of generality that $|f|$ is bounded by 1. The function $g(y,z) = E(f(X) \mid Y = y, Z = z)$, is continuous in $(y,z) \in \mathrm{Supp}(Y,Z)$ which is a closed set. By the classical Tietze-Urysohn extension theorem it can be extended to a function continuous everywhere and still taking values in $[-1,1]$. In the rest of the proof when we write $E(f(X) \mid Y = y, Z = z)$ we will always mean this continuous extension.

Without loss of generality restrict z, z_0 to a compact set of values of z, and choose a compact set K of values of y such $\liminf_{z \to z_0} \Pr(Y \in K \mid Z = z) > 1 - \varepsilon$ where ε is arbitrarily small. Write

$$\begin{aligned}
\int & E(f(X) \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z) \\
&= \int_{y \in K} E(f(X) \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z) \\
&+ \int_{y \notin K} E(f(X) \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z).
\end{aligned}$$

The second term on the right-hand side is smaller than ε for z close enough to z_0 (and for $z = z_0$). In the first term on the right-hand side, the integrand $E(f(X) \mid Y = y, Z = z)$ is a continuous function of (y,z), which varies in a product of two compact sets. It is therefore uniformly continuous in (y,z), and hence continuous in z, uniformly in y. Therefore for z close enough to z_0, $\int E(f(X) \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z)$ is within 2ε of $\int_K E(f(X) \mid Y = y, Z = z_0) \Pr(Y \in \mathrm{d}y \mid Z = z)$. Again for z close enough to z_0, this is within 3ε of $\int E(f(X) \mid Y = y, Z = z_0) \Pr(Y \in \mathrm{d}y \mid Z = z)$. Since the integrand here is a fixed bounded continuous function of y, for $z \to z_0$ this converges to $\int E(f(X) \mid Y = y, Z = z_0) \Pr(Y \in \mathrm{d}y \mid Z = z_0)$. Thus for z close enough to z_0, $\int E(f(X) \mid Y = y, Z = z) \Pr(Y \in \mathrm{d}y \mid Z = z)$ is within 4ε of $\int E(f(X) \mid Y = y, Z = z_0) \Pr(Y \in \mathrm{d}y \mid Z = z_0)$. $\square$

3. The g-computation formula for continuous variables.

We will solve the uniqueness problem before tackling the more difficult correctness issue. First we present a natural generalisation of condition A3:

A3*: Evaluability. For any $\overline{a}_k = g(\overline{l}_k)$ and $(\overline{l}_k, \overline{a}_{k-1}) \in \mathrm{Supp}((\overline{L}_k, \overline{A}_{k-1}))$, it follows that $(\overline{l}_k, \overline{a}_k) \in \mathrm{Supp}((\overline{L}_k, \overline{A}_k))$.

As with the original version of A3, the condition calls a plan g evaluable if, whenever at some stage there was an opportunity to use the plan, it was indeed implemented on some proportion of the patients. If all variables are actually discrete then A3* reduces to the original A3.

Next we summarize appropriate continuity conditions concerning the factual variables.

C: Continuity. The distributions $\mathrm{law}(Y \mid \overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K)$ can be chosen continuous in $(\overline{l}_K, \overline{a}_K)$, and $\mathrm{law}(L_k \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1})$ in $(\overline{l}_{k-1}, \overline{a}_{k-1})$, on the (joint) supports of the conditioning variables.

Theorem 1. *Suppose conditions A3* and C hold. Then the right-hand side of (1) is unique when the conditional distributions on the right-hand side are chosen subject to continuity.*

Proof. The right-hand of (1) has the probabilistic interpretation that first a value l_1 is generated according to $\mathrm{law}(L_1)$, then a_1 is specified by $a_1 = g_1(l_1)$, then a value l_2 is generated from $\mathrm{law}(L_2 \mid L_1 = l_1, A_1 = a_1)$, and so on. Suppose that at the end of the kth step we have obtained $(\bar{l}_k, \bar{a}_k) \in \mathrm{Supp}(\overline{L}_k, \overline{A}_k)$. Then l_{k+1} will with probability one be generated, according to a uniquely determined probability distribution, in $\mathrm{Supp}(L_{k+1} \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k)$, thus $(\bar{l}_{k+1}, \bar{a}_k) \in \mathrm{Supp}(\overline{L}_{k+1}, \overline{A}_k)$ by Fact 3. By condition A3*, this leads to $(\bar{l}_{k+1}, \bar{a}_{k+1}) \in \mathrm{Supp}(\overline{L}_{k+1}, \overline{A}_{k+1})$. By induction, with probability one all values of l_k (and in the last step, of y), are generated from uniquely determined conditional distributions. □

We now have conditions under which the functional $b(g; \mathrm{law}(\overline{L}, \overline{A}, Y))$ on the right-hand side of (1) is well-defined. We next want to investigate when it equals $\mathrm{law}(Y^g)$. For that we need supplementary continuity conditions on its conditional laws given the factual variables, and then appropriately reformulated versions of assumptions A1 and A2. We first state suitable supplementary continuity conditions Cg.

Cg: Continuity for counterfactuals. The distributions $\mathrm{law}(Y^g \mid \overline{L}_{k+1}, \overline{A}_k)$ and $\mathrm{law}(Y^g \mid \overline{L}_k, \overline{A}_k)$ can for all k all be chosen continuous in the values of the conditional variables on their supports.

Continuity assumptions C and Cg imply that conditional distributions selected according to continuity are uniquely defined on the relevant supports. In the the sequel, in particular in the following alternative versions of assumptions A1 and A2, all conditional distributions are taken to be precisely those prescribed by continuity:

A1*: Consistency. $\mathrm{law}(Y^g \mid \overline{L} = \bar{l}, \overline{A} = \bar{a}) = \mathrm{law}(Y \mid \overline{L} = \bar{l}, \overline{A} = \bar{a})$ for $(\bar{l}, \bar{a}) \in \mathrm{Supp}(\overline{L}, \overline{A})$ and $g(\bar{l}) = \bar{a}$.

A2*: Randomisation. $\mathrm{law}(Y^g \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k)$ does not depend on a_k for $\bar{a}_k, \bar{l}_k \in \mathrm{Supp}(\overline{L}_k, \overline{A}_k)$ and satisfying $\bar{a}_{k-1} = g(\bar{l}_{k-1})$.

Theorem 2. *Suppose conditions C and Cg hold, and moreover assumptions A1*–A3* hold. Then equation (1) is true.*

Proof. Writing out A1*, we have that

$$\Pr(Y^g \in \cdot \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K) = \Pr(Y \in \cdot \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K) \tag{3}$$

for $(\bar{l}_K, \bar{a}_K) \in \mathrm{Supp}(\overline{L}_K, \overline{A}_K)$ and $g(\bar{l}_K) = \bar{a}_K$, where both conditional distributions are uniquely determined by continuity. Now let $(\bar{l}_{k-1}, \bar{a}_{k-1}) \in \mathrm{Supp}(\overline{L}_{k-1}, \overline{A}_{k-1})$ and satisfying $g(\bar{l}_{k-1}) = \bar{a}_{k-1}$ be fixed. Consider

$$\int_{l_k \in \mathrm{Supp}(L_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}); a_k = g_k(\bar{l}_k)} \begin{array}{l} \Pr(Y^g \in \cdot \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k) \\[2ex] \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}). \end{array} \tag{4}$$

Since $l_k \in \mathrm{Supp}(L_k \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1})$ we have $(\overline{l}_k, \overline{a}_{k-1}) \in \mathrm{Supp}(\overline{L}_k, \overline{A}_{k-1})$ by Fact 3. By assumption A3* and Fact 3 again, this gives us $(\overline{l}_k, \overline{a}_k) \in \mathrm{Supp}(\overline{L}_k, \overline{A}_k)$. Hence all conditional distributions in (4) are well defined. By A2* we can delete the condition $A_k = a_k$ in $\Pr(Y^g \in \cdot \mid \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k)$. The integrand now does not depend on a_k and integrating out l_k shows that (4) is equal to *a version of*

$$\Pr(Y^g \in \cdot \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}). \tag{5}$$

However it is not obvious, that this is the same version indicated by continuity. Fact 4 however states that continuously mixing over one parameter, a family of distibutions continuous in two parameters, results in a continuous family. Consequently (5) *is* the version selected by continuity.

The theorem is now proved exactly as in the discrete case by repeating the step which led from (4) to (5) for $k = K, K-1, \ldots, 1$ on the right hand side of (1) (after replacing Y by Y^g), at the end of which the left hand side of (1) results. □

In view of Fact 4, the continuity condition Cg would be a lot more simple if we could assume not only, from condition C, that $\mathrm{law}(L_k \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1})$ is continuous in $(\overline{l}_{k-1}, \overline{a}_{k-1})$, but also

Ca: Continuity of factual treatment distribution. $\mathrm{law}(A_k \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1})$ is continuous in $(\overline{l}_k, \overline{a}_{k-1})$.

Then for Cg it suffices to assume that $\mathrm{law}(Y^g \mid \overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K)$ is continuous in $(\overline{l}_K, \overline{a}_K)$ since by mixing it alternately with respect to the conditional laws of A_k and L_k, $k = K, K-1, \ldots, 1$ maintains at each stage, according to Fact 4 with Ca and Cg respectively, the continuity in the remaining conditioning variables.

When the covariates and treatments are discrete condition A2* reduces to the original A2. Assumption A1* on the other hand is then weaker than A1. One might prefer stronger continuity assumptions and a stronger version of A1* which would reduce to A1 with discrete variables; for instance assume that $\mathrm{law}((Y, Y^g) \mid \overline{L}, \overline{A})$ can be chosen continuous in the conditioning variables on their support, and assume that with respect to this version, $\Pr(Y = Y^g \mid \overline{L} = \overline{l}, \overline{A} = \overline{a}) = 1$ for $\overline{a} = g(\overline{l})$. Informally this says that Y and Y^g coincide with larger and larger probability, the closer the plan g has been adhered to.

It would be interesting to show, without any continuity assumptions at all, that the g-computation formula is correct for almost all plans g, where we have to agree on an appropriate measure on the space $\mathcal{G}$ of all plans g. So far we were not able to settle this question. It arises again in when we consider the alternative approach based on randomised plans in section 6.

4. Characterizing the null-hypothesis.

The g-computation formula plays a major role in the Robins' (1986, 1987, 1989, 1997) theory of causal inference for complex longitudinal data, through the proofs of some theorems giving necessary and sufficient conditions for the "g"-null hypothesis H_0 that the right hand side of (1) is the same for all evaluable treatment plans g; by this we now mean a plan g satisfying the evaluability assumption A3*. These theorems concern various functionals of the distribution of the factual variables only. We will therefore only assume the continuity conditions C. Under the further conditions making (1) not only unique but also correct, the "g"-null hypothesis is equivalent to the more interesting g-null hypothesis that the distribution of the outcome under any evaluable plan g is the same, and hence treatment indeed has no effect on outcome.

We call a treatment plan static if it does not depend in any way on the covariate values $\overline{l}$, in other words, it is just a fixed sequence of treatment values $a_1, \ldots, a_K$ to be assigned at each time point irrespective of covariate values measured then or previously. A dynamic plan is just a plan which is not static.

Some of the results use the concept of a baseline treatment plan. In the literature this has been usually taken to be the static plan $g \equiv \overline{0} = (0, \ldots, 0)$ where 0 is a special value in each A_k's sample space. However, already in the discrete case, complications arise if this plan, and plans built up from another plan g by switching from some time point from the plan g to the plan $\overline{0}$, are not evaluable. (Thanks to Judith Lok for bringing this to our attention).

We will say that a plan g^0 is an admissible baseline plan if for all evaluable plans g and all $k = 0, \ldots, K$, the plan $g^{k:0}$ (follow plan g up to and including time point $k-1$, follow plan g_0 from time point k onwards) is also evaluable. We assume that an admissible baseline plan exists. It is possible to construct examples where none exists; and certainly easy to construct examples where no static admissible baseline plan exists. The problem is that even if x is a support point of the law of a random variable X, there need not exist any y such that (x, y) is a support point of the law of (X, Y). Admissible baseline plans exist if condition Ca holds, by appeal to Fact 3; and they exist if the sample space for each treatment is compact.

For a given plan g, for given k, and given $(\overline{l}_k, \overline{a}_{k-1})$, introduce the quantity

$$b(g; \overline{l}_k, \overline{a}_{k-1}) = \int\limits_{l_{k+1}} \ldots \int\limits_{l_K} \Pr(Y \in \cdot \mid \overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K) \prod_{k'=k+1}^{K} \Pr(L_{k'} \in \mathrm{d}l_{k'} \mid \overline{L}_{k'-1} = \overline{l}_{k'-1}, \overline{A}_{k'-1} = \overline{a}_{k'-1}) \tag{6}$$

where $a_k, \ldots, a_K$ on the right hand side are taken equal to $g_k(\overline{l}_k), \ldots, g_K(\overline{l}_K)$. Similarly to Theorem 1, this is a well-defined functional of the joint law of the factual variables when $(\overline{l}_k, \overline{a}_{k-1})$ lies in the support of $(\overline{L}_k, \overline{A}_{k-1})$, when $g(\overline{l}_{k-1}) = \overline{a}_{k-1}$, and when g is evaluable, if conditional distributions are chosen subject to continuity in distribution on the support of the conditioning variables. In fact the expression (6) does not depend on g at time points prior to the k'th, so it is well-defined more generally than this. Let us say that a plan g is k-evaluable relatively to a given $(\bar{l}_k, \bar{a}_{k-1})$ if for all $m \geq k$, any $(\bar{l}_m, \bar{a}_{m-1}) \in \text{Supp}(\overline{L}_m, \overline{A}_{m-1})$ with initial segments coinciding with $\bar{l}_k$ and $\bar{a}_{k-1}$ and satisfying $g_j(\bar{l}_j) = a_j$ for $j = k, \ldots, m-1$, we have $(\bar{l}_m, \bar{a}_m) \in \text{Supp}(\overline{L}_m, \overline{A}_m)$ where of course $g_m(\bar{l}_m) = a_m$.

Similarly to Theorem 2 one has under appropriate conditions that $b(g; \bar{l}_k, \bar{a}_{k-1}) = \text{law}(Y^g \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1})$, but this interpretation plays no role in the sequel.

The theorems we want to prove are the following:

Theorem 3. *Assume condition C. Under H_0, for any k and $(\bar{l}_k, \bar{a}_{k-1})$ in the support of $(\overline{L}_k, \overline{A}_{k-1})$, the expression $b(g; \bar{l}_k, \bar{a}_{k-1})$ does not depend on g for any k-evaluable plans g.*

Theorem 4. *Assume condition C. Suppose an admissible baseline plan g^0 exists. Then if $b(g^{k+1:0}; \bar{l}_k, \bar{a}_{k-1})$ does not depend on $a_k = g_k(\bar{l}_k)$ for all $(\bar{l}_k, \bar{a}_k)$ in the support of $(\overline{L}_k, \overline{A}_k)$, then H_0 is true.*

Note in Theorem 4 that $b(g^{k+1:0}; \bar{l}_k, \bar{a}_{k-1})$ *only* depends on g through the value a_k of $g_k(\bar{l}_k)$.

Theorem 5. *Assume condition C. Suppose an admissible baseline plan g^0 exists. Then H_0 holds if and only if $Y \perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$ for all k.*

Theorem 6. *Assume condition C and suppose an admissible baseline plan g^0 exists. Suppose the functions $\gamma_k = \gamma_k(y; \bar{l}_k, \bar{a}_k)$ can be found satisfying the following: if a random variable Y^k has the distribution $b(g^{k+1:0}; \bar{l}_k, \bar{a}_{k-1})$ where $g_k(\bar{l}_k) = a_k$ then $\gamma_k(Y^k; \bar{l}_k, \bar{a}_k))$ is distributed as $b(g^{k:0}; \bar{l}_k, \bar{a}_{k-1})$. Define $Y^K = Y$ and then recursively define $Y^{k-1} = \gamma_k(Y^k; \overline{L}_k, \overline{A}_k)$. Then Y^0 satisfies $Y^0 \perp A_k | \overline{L}_k, \overline{A}_k$, for all k.*

If Y is real-valued and continuously distributed then the obvious choice for the functions γ_k in Theorem 6 is the QQ-transform between the specified distributions.

Combining theorems 3 and 4 we obtain two further 'if and only if' results: assuming condition C and that an admissible baseline plan g^0 exists, H_0 is true if and only if $b(g; \bar{l}_k, \bar{a}_{k-1})$ does not depend on g for any k-evaluable plans g, and if and only if $b(g; \bar{l}_k, \bar{a}_{k-1})$ does not depend on g for any plan of the special form $g^{k+1:0}$. In particular, if $g_0 \equiv \overline{0}$ is an evaluable baseline plan, then H_0 holds if and only if $b(g; \bar{l}_k, \bar{a}_{k-1})$ does not depend on g for any *static* plan g.

Theorem 5 shows that testing of the null-hypothesis does not require one to actually estimate and compute (1) for all plans g, and resolves the problem that, were one to estimate the component conditional distributions of (1) using parametric models (non-saturated), then typically no combination of parameter values could even reproduce the null-hypothesis (Robins, 1997; Robins and Wasserman, 1997). Theorem 4 and Theorem 6 are the starting point of a new parametrization in which one models the effect $\gamma_k(y; \bar{l}_k, \bar{a}_k)$ of one final 'blip' of treatment a_k at time-point k before reverting to the base-line treatment g^0. Parametric models for these effects, which Robins (1989, 1997) refers to as structural nested models, do enable one to cover the null-hypothesis in a simple way and lead to estimation and testing procedures which are mututally consistent and robust to misspecification, at least, at the null hypothesis. Briefly, the variable Y^0 constructed in Theorem 6 can be used as a surrogate for Y^{g^0}. One can estimate parameters of the blip-down functions γ_k by testing the hypotheses that $Y^0 \perp A_k | \overline{L}_k, \overline{A}_k$. This method of estimation is discussed in detail in Robins (1997) under the rubric of g-estimation of structural nested models.

Proof of Theorem 3. Suppose H_0 is true. Consider two plans g^1 and g^2. We want to prove equality of $b(g^i;\overline{l}^0_k,\overline{a}^0_{k-1})$ for $i=1,2$, where the superscript 0 is used to distinguish the fixed values given in the theorem from later variable ones. Since b does not depend on either plan g^i before time k, without loss of generality suppose that these two plans assign treatments a^0_1, ..., a^0_{k-1} statically over the first $k-1$ time-points. Fix $\varepsilon>0$ and define the plan g^3 to be identical to plan g^1 except that for $m\geq k$ and $\overline{l}_m$ for which $\overline{l}_k$ is in an epsilon ball about $\overline{l}^0_k$, it is identical to g_2. Consider the equality of the two probability distributions $b(g^1)$ and $b(g^3)$ on any given event in the sample space for Y. As we integrate over all $l_1,\ldots,l_K$ we are integrating identical integrands except for $\overline{l}_k$ in the epsilon ball about $\overline{l}^0_k$ which is precisely where g^1 and g^3 differ; denote this set $B(\overline{l}^0_k,\varepsilon)$. Deleting the integrals over the complement of this set we obtain the equality, for $i=1,2$, of the two quantities

$$\int_{\overline{l}_k\in B(\overline{l}^0_k,\varepsilon)} b(g^i;\overline{l}_k,\overline{a}^0_{k-1})\prod_1^k \Pr(L_j\in \mathrm{d}l_j \mid \overline{L}_{j-1}=\overline{l}_{j-1},\overline{A}_{j-1}=\overline{a}^0_{j-1}). \tag{7}$$

Now by our continuity assumptions and repeated use of Fact 4, $b(g^i;\overline{l}_k,\overline{a}^0_{k-1})$ is a continuous function of $\overline{l}_k$. Divide (7) by the normalising quantity $\int_{\overline{l}_k\in B(\overline{l}^0_k,\varepsilon)}\prod_1^k \Pr(L_j\in \mathrm{d}l_j \mid \overline{L}_{j-1}=\overline{l}_{j-1},\overline{A}_{j-1}=\overline{a}^0_{j-1})$; the same for both $i=1,2$. Now the equality expresses the equality of the expectations of $b(g_i;\overline{L}^\varepsilon_k,\overline{a}^0_{k-1})$ for $i=1,2$ where $\overline{L}^\varepsilon_k$ lies with probability one in $B(\overline{l}^0_k,\varepsilon)$. As $\varepsilon\to 0$, by continuity of $b(g_i;\cdot,\overline{a}^0_{k-1})$, the expectations converge to $b(g_i;\overline{l}^0_k,\overline{a}^0_{k-1})$. $\square$

Proof of Theorem 4. Let g be a given evaluable plan. Recall that $g^{k:0}$ denotes the modification of the plan obtained by making all treatments from time k onward follow the baseline plan g^0. Let $g^{k:a_k,0}$ denote the modification of the given plan g obtained by making the k'th treatment equal to the fixed amount a_k and all subsequent treatments follow the baseline plan. We show by downwards induction on k that $b(g;\overline{l}_k,\overline{a}_{k-1})=b(g^{k:0};\overline{l}_k,\overline{a}_{k-1})$ for all k. This statement for $k=0$ is the required conclusion. To initialise the induction note that $b(g;\overline{l}_K,\overline{a}_{K-1})=b(g^{K+1:0};\overline{l}_K,\overline{a}_{K-1})=b(g^{K:0};\overline{l}_K,\overline{a}_{K-1})$, where the first equality is trivial and the second is the assumption of the theorem for $k=K$. Next, in general, write

$$
\begin{aligned}
b(g;\bar{l}_k,\bar{a}_{k-1}) &= \int_{l_{k+1}} b(g;\bar{l}_{k+1},\bar{a}_k)\,\Pr(L_{k+1}\in \mathrm{d}l_{k+1}\mid \overline{L}_k=\bar{l}_k,\overline{A}_k=\bar{a}_k)\\
&= \int_{l_{k+1}} b(g^{k+1:0};\bar{l}_{k+1},\bar{a}_k)\,\Pr(L_{k+1}\in \mathrm{d}l_{k+1}\mid \overline{L}_k=\bar{l}_k,\overline{A}_k=\bar{a}_k)\\
&\qquad\qquad \text{(by the induction hypothesis)}\\
&= b(g^{k+1:0};\bar{l}_k,\bar{a}_{k-1})\\
&= b(g^{k:g_k(\bar{l}_k),0};\bar{l}_k,\bar{a}_{k-1})\\
&\qquad\qquad \text{(by inspection)}\\
&= b(g^{k:0};\bar{l}_k,\bar{a}_{k-1})\\
&\qquad\qquad \text{(by the assumption of the theorem)}
\end{aligned}
$$

which establishes the induction step. □

Proof of Theorem 5. We prove first the backwards implication. Given that $Y \perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$ we see that Y itself satisfies the assumptions Cg, A1* and A2* concerning Y^g, for any particular evaluable g, of Theorem 2. Thus its law is given by the g-computation formula (1) which is therefore the same for all g.

For the forward implication, we show that $Y \not\perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$ for some k implies the existence of some k and evaluable plans g for which $b(g;\bar{l}_k,\bar{a}_{k-1})$ depends on g. First of all, note there must be a *last* k, say $k=k_0$, for which the conditional independence does not hold. Now in the g-computation formula (1), for $k=K,K-1,\ldots,k_0+1$ we can repeatedly a) drop the last a_k in the integrand, by conditional independence, and b) integrate out the last l_k. Thus the g-computation formula holds with K replaced by k_0, and we can replace K by k_0 in all subsequent results. But now we see by inspection that $b(g;\bar{l}_{k_0},\bar{a}_{k_0-1})$, which is nothing but the conditional law of Y given $\overline{L}_{k_0},\overline{A}_{k_0}$, depends on $a_{k_0}=g_{k_0}(\bar{l}_{k_0})$ and by Theorem 5 we are done. □

Proof of Theorem 6. By downwards induction one verifies that for each k, Y^k has the conditional distribution $b(g^{k+1:0};\bar{l}_k,\bar{a}_{k-1}$ given $\overline{L}_k=\bar{l}_k,\overline{A}_k=\bar{a}_k$, where $g_k(\bar{l}_k)=a_k$. Given $(\overline{L}_k,\overline{A}_{k-1})$, Y^0 is a deterministic function of $Y^{k-1}=\gamma_k(Y^k;\overline{L}_k,\overline{A}_k)$. So it suffices to verify that $\gamma_k(Y^k;\overline{L}_k,\overline{A}_k)\perp A_k|\overline{L}_k,\overline{A}_{k-1}$. This follows by the characterizing property of γ_k and the just stated conditional distribution of Y^k. □

5. Construction of counterfactuals.

Suppose we start with a given law$(\overline{L}, \overline{A}, Y)$. Can we build on a new sample space the same random variables (i.e., variables with the same joint distribution) together with counterfactuals Y^g for all g, satisfying conditions A1–A3 (or their strengthened versions)? The answer will be yes. This means that in whatever sense counterfactuals exist or do not exist, it is harmless to pretend that they do exist and to investigate the consequences of that assumption—we do not hereby impose 'hidden' restrictions on the distribution of the data.

The solution we give to this problem works in the reverse direction: we construct the counterfactual world first, then build the factual world on top of it. However once we have constructed all variables together with the required properties, including the factuals with their given distribution, we can now derive the conditional distribution of all counterfactuals given all factuals, and hence we can extend a sample space supporting just the factual variables with all the counterfactuals as well, just by using auxiliary randomization.

Fix a collection of versions of laws of each L_k, A_k and Y given all their predecessors (in the usual order L_1, A_1, $\ldots$, L_K, A_K, Y). A plan g_0 is called static if it does not depend on $\overline{l}$; i.e., it is just a single sequence of treatments a_k to be applied irrespective of the measured covariate values. Let $\mathcal{G}_0$ denote the collection of static plans; it can be identified with the collection of all $\overline{a}$.

First we build random variables $\overline{L}^{g_0}$, Y^{g_0} for all $g_0 \in \mathcal{G}_0$. Generate L_1 from its marginal law. For all g_0, $L_1^{g_0} = L_1$. Next, for each value of a_1 generate a random variable $L_2^{l_1,a_1}$ from the law of L_2 given $L_1 = l_1, A_1 = a_1$. For all g_0 with $(g_0)_1 = a_1$, define $L_2^{g_0} = L_2^{l_1,a_1}$ on $L_1^{g_0} = l_1$. Proceed in the same way finishing with a collection of variables $Y^{l_1,a_1,\ldots,l_K,a_K}$ drawn from the laws of Y given $\overline{L} = \overline{l}, \overline{A} = \overline{a}$ and define $Y^{g_0} = Y^{l_1,a_1,\ldots,l_K,a_K}$ on $L_1^{g_0} = l_1$, $\ldots$, $L_K^{g_0} = l_K$; $(g_0)_1 = a_1$, $\ldots$, $(g_0)_K = a_K$. Note that the definition of $L_k^{g_0}$ only depends on the values of $(g_0)_1, \ldots, (g_0)_{k-1}$.

For definiteness, we could use at each stage a single independent uniform-$[0, 1]$ variable U_k to generate all $L_k^{g_0}$.

Now we can define counterfactuals Y^g, L_k^g for the dynamic plans g by using the recursive consistency rule: $L_k^g = L_k^{g_0}$ where $(g_0)_{k-1} = g_{k-1}(\overline{L}_{k-1}^g)$, and similarly $Y^g = Y^{g_0}$ where $(g_0)_K = g_K(\overline{L}_K^g)$. Note that when for instance we set $L_k^g = L_k^{g_0}$, values of $(g_0)_1, \ldots, (g_0)_{k-2}$ have already been determined and only the next value $(g_0)_{k-1}$ is still unknown, for which we use the rule $(g_0)_{k-1} = g_{k-1}(\overline{L}_{k-1}^g)$.

On top of the counterfactual world we now define the 'real world', the factuals $\overline{L}, \overline{A}, Y$. To build these variables we use a new sequence of independent uniform random variables successively as follows: $L_k = L_k^{g_0}$ where $(g_0)_{k-1} = A_{k-1}$; A_k is drawn from the prespecified law of A_k given $\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}$ on the event $\overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}$. Finally $Y = Y^{g_0}$ where $(g_0)_K = A_K$. As before successive values of g_0 are generated as they are needed. One should check that the resulting $\overline{L}, \overline{A}, Y$ do indeed have the intended joint distribution.

The consistency assumption A1 holds by construction. The randomisation assumption A2 holds in the very strong form $(Y^g : g \in \mathcal{G}) \perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$ where $\mathcal{G}$ is the set of all treatment plans. This follows since given all Y^g and given $(\overline{L}_k, \overline{A}_{k-1})$, we used a single independent uniform $[0, 1]$ variable and the values of $(\overline{L}_k, \overline{A}_{k-1})$ only in order to construct A_k. Whether or not the evaluability condition A3 holds depends of course on which plan g

is being considered. The collection of conditional distributions we used to start with is not uniquely defined in the continuous case, and also not uniquely defined in the discrete case if not all values of $\overline{L}, \overline{A}$ have positive probability. However under the continuity conditions C, if we have chosen all conditional distributions subject to continuity on the supports of the conditioning variables, then our construction satisfies the stronger conditions Cg, A1* and A2*.

6. A G-computation formula for randomised plans.

In this section we present an alternative solution to the problems posed at the beginning of the paper. Instead of assuming continuity of conditional distributions, is to assume a kind of continuity of the treatment plan g relative to the factual plan. Our problems before arose because the deterministic plan g was not actually implemented with positive probability, when covariates are continuously distributed. Suppose we allow plans by which the amount of treatment allocated at stage k, given the past, has some random variation. In practice this actually is the often the case, for instance, it may be impossible to exactly deliver a certain amount of a drug, or to exactly measure a covariate. Note that in the theory below the variables A_k and L_k are the actually administered drug quantity, and the true value of the covariate; thus from a statistical point of view our theory may not be of direct use since these variables will in practice not be observed. Imagine that all variables are measured precisely and random treatments can be given according to any desired probability distribution.

A randomised treatment plan now denoted by G consists of a sequence of conditional laws $\Pr(A_k^G \in \cdot \mid \overline{L}_k^G = \overline{l}_k, \overline{A}_{k-1}^G = \overline{a}_{k-1})$. (The random variables $A_k^G, \overline{L}_k^G$ and $\overline{A}_{k-1}^G$ here are counterfactuals corresponding to plan G being adhered to from the start).

The G-computation formula now becomes

$$\begin{aligned}\Pr(Y^G \in \mathrm{d}y) = \int_{l_1}\int_{a_1}\cdots\int_{l_K}\int_{a_K} & \Pr(Y \in \mathrm{d}y \mid \overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K) \\ & \prod_{k=1}^{K} \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1})\cdot \\ & \cdot \Pr(A_k^G \in \mathrm{d}a_k \mid \overline{L}_k^G = \overline{l}_k, \overline{A}_{k-1}^G = \overline{a}_{k-1}).\end{aligned} \tag{8}$$

Again questions of uniqueness and correctness arise. Uniqueness of the right-hand side of (8), denoted $b(G; \mathrm{law}(\overline{L}, \overline{A}, Y))$ is easy to check under the following generalization of assumption A3:

A3: Evaluability.** For each k, $\mathrm{law}(A_k^G \mid \overline{L}_k^G = \overline{l}_k, \overline{A}_{k-1}^G = \overline{a}_{k-1})$ is absolutely continuous with respect to $\mathrm{law}(A_k \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1})$ for almost all $(\overline{l}_k, \overline{a}_{k-1})$ from the law of $\overline{L}_k, \overline{A}_{k-1}$.

Theorem 6. *Under A3**, $b(G; \mathrm{law}(\overline{L}, \overline{A}, Y))$ is uniquely defined by the right-hand side of (8).*

Proof. Consider the expression

$$\begin{aligned} &\int_{l_1}\int_{a_1}\cdots\int_{l_K}\int_{a_K} \Pr(Y \in \mathrm{d}y \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \overline{a}_K)\cdot \\ \prod_{k=1}^{K} &\frac{\mathrm{dP}_{A_k^G \mid \overline{L}_k^G = \bar{l}_k, \overline{A}_{k-1}^G = \overline{a}_{k-1}}}{\mathrm{dP}_{A_k \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}}} \cdot \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1})\cdot \\ &\cdot \Pr(A_k \in \mathrm{d}a_k \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}). \end{aligned} \tag{9}$$

The successive integrations with repect to the conditional laws of L_k and A_k could be rewritten as a single integration with respect to the joint law of $(\overline{L}_K, \overline{A}_K)$. Moreover (9) does not depend on choice of Radon-Nikodym derivatives nor on choice of the conditional law of Y, since all are almost surely unique and by A3** finite on the support of $\overline{L}_K, \overline{A}_K$. Now in (9) we can successively, for $k = K, K-1, \ldots, 1$ merge the kth Radon-Nikodym derivative and integration with respect to the conditional law of A_k, replacing it by integration with respect to the conditional law of A_k^G. This transforms (9) into the right-hand side of (8), showing that (8) too does not depend on choice of Radon-Nikodym derivatives or conditional distributions. □

Condition A3** can be weakened; we only need the absolute continuity along paths $\bar{l}_K, \overline{a}_K$ which can actually be realised.

Does (8) also give the correct answer? This requires introducing a counterfactual Y^G and relating it to Y^g and Y.

Suppose a plan G is to be implemented by, at each stage, generating A_k^G from the specified conditional law by a transformation of an independent uniform variable U_k. We could generate the U_k in advance, and thereby generate a candidate A_k^G for all possible intermediate values of $(\overline{L}_k^G, \overline{A}_{k-1}^G)$; call it $a_k^G(\bar{l}_k, \overline{a}_{k-1}; u_k)$. Tracking through all possible values of all L_k^G, we see that the randomised plan G is exactly equivalent to choosing in advance, by a randomisation depending only on $U_1, \ldots, U_K$, a non-randomized plan $g = g_{\overline{u}}$. A little thought shows that the right-hand side of (6) can be rewritten as $\int \ldots \int b(g_{\overline{u}}; \mathrm{law}(\overline{L}, \overline{A}, Y))\mathrm{d}u_1 \ldots \mathrm{d}u_K$. So if we make the additional consistency assumption $Y^G = Y^g$ on $G = g$, then (8) gives a *correct* expression for $\mathrm{law}(Y^G)$ as long as (1) is correct for all (or at least, almost all) g.

Now we know already that the right-hand side of (8) is unique. So if versions of all conditional laws *could* be chosen simultaneously making (1) correct for almost all g, then by taking those choices, and averaging (1) over g, produces not only the unique but also the correct expression (8). However it is not clear if this can be done.

If we are going to make assumptions concerning all Y^g simultaneously, other routes become available. Rather than working via (1) for each g separately, we can try directly to establish (8). But in order to be able to work with joint conditional laws of all Y^g simultaneously, we have to assume a lot of regularity. We will do it here by assuming that the probability space on which all random variables are defined is nice enough (one could say, small enough), that conditional probability measures or so-called disintegrations (see Chang and Pollard, 1997) over this space exist. This will have the further advantage that we can once and for all choose versions of all conditional probability measures in a mutually consistent way; we automatically obtain the correct version of a given conditional probability measure when mixing over one of the conditioning variables.

A0: Sample space regularity**. The underlying probability space $(\Omega, \mathcal{F}, \Pr)$ is a complete separable metric space with the Borel σ-algebra.

Fix a disintegration of Pr with respect to L_1, then fix disintegrations of $\Pr(\cdot \mid L_1 = l_1)$ with respect to A_1, and so on. We now have, everywhere on Ω,

$$\begin{aligned}\int_{a_k} \Pr(\cdot \mid \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k) \Pr(A_k \in \mathrm{d}a_k \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) \\ = \Pr(\cdot \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1})\end{aligned}$$

and similarly

$$\begin{aligned}\int_{l_k} \Pr(\cdot \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}) \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}) \\ = \Pr(\cdot \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}).\end{aligned}$$

The conditional probability measures here are measures on Ω, concentrated on the conditioning event.

We are going to talk about conditional joint laws of all Y^g simultaneously; denoting by $\mathcal{G}$ the set of all plans g let $Y^{\mathcal{G}}$ denote this collection of random variables. By its law or conditional law we mean the restriction of Pr or appropriate conditional distribution, to the sub-σ-algebra of $\mathcal{F}$ generated by all Y^g.

Consider the following versions of A1 and A2.

A1: Consistency.** $Y^G = Y^g$ on $G = g$ and, for each g, $Y^g = Y$ on $g(\overline{L}) = \overline{A}$.

A2: Randomisation.** $Y^{\mathcal{G}} \perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$.

Theorem 7. *Under A0**–A3**, formula (8) is correct.*

Proof. By A2**, for almost all $\overline{l}_k, \overline{a}_{k-1}$, $\mathrm{law}(Y^{\mathcal{G}} \mid \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k)$ does not depend on a_k, for almost all a_k with respect to $\Pr(A_k \in \cdot \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1})$. So by mixing over A_k from its conditional law, we find that $\mathrm{law}(Y^{\mathcal{G}} \mid \overline{L}_k = \overline{l}_k, \overline{A}_k = \overline{a}_k)$ coincides with $\mathrm{law}(Y^{\mathcal{G}} \mid \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1})$ for almost all $\overline{l}_k, \overline{a}_k$.

These 'almost all' statements refer to the factual law of $\overline{L}, \overline{A}$, but by A3** they also hold almost everywhere with respect to the integrating measure in (8). Now (8) can be rewritten as

$$\begin{aligned}\int_{u_1} \cdots \int_{u_K} \int_{l_1} \cdots \int_{l_K} \Pr(Y \in \cdot \mid \overline{L} = \overline{l}, \overline{A} = \overline{a}) \\ \cdot \prod_{k=1}^{K} \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \overline{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1}) \\ \cdot \mathrm{d}u_1 \ldots \mathrm{d}u_K\end{aligned} \tag{10}$$

where $a_k = a_k^G(\bar{l}_k, \bar{a}_{k-1}; u_k)$, $k = 1, \ldots, K$. We can successively simplify (10) as follows. First, by A1** we can replace Y by Y^g where $g = g_{\overline{u}}$. Here we use the fact that we have disintegrations, so that if $Y = Y^g$ on a certain event the conditional laws of these variables are the same given this same event. Next by A2** for $k = K$, we can delete the conditioning $A_K = a_K$ in $\Pr(Y^g \in \cdot \mid \overline{L} = \bar{l}, \overline{A} = \overline{a})$, at least, for almost all $\bar{l}$, $\overline{a}$. The exceptions do not however change the value of the integral. Moreover we can do this irrespective of the value of $g = g_{\overline{u}}$. Now we may mix over the conditional law of L_K, reducing (10) to

$$\int_{u_1} \cdots \int_{u_K} \int_{l_1} \cdots \int_{l_{K-1}} \Pr(Y^g \in \cdot \mid \overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \overline{a}_{K-1})$$
$$\cdot \prod_{k=1}^{K-1} \Pr(L_k \in \mathrm{d}l_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \overline{a}_{k-1})$$
$$\cdot \mathrm{d}u_1 \ldots \mathrm{d}u_K$$

where $a_k = a_k^G(\bar{l}_k, \overline{a}_{k-1}; u_k)$ $k = 1, \ldots, K-1$ and $g = g_{\overline{u}}$. Repeat a further $K-1$ times and we finally obtain

$$\int_{u_1} \cdots \int_{u_K} \Pr(Y^{g_{\overline{u}}} \in \cdot) \mathrm{d}u_1 \ldots \mathrm{d}u_K = \Pr(Y^G \in \cdot).$$

□

The above theory is not just a distributional theory. We have assumed specific facts about the underlying sample space, involving events of zero probability. In particular the consistency assumption is back in its original form for discrete variables.

Acknowledgements.

Richard Gill is grateful to the Department of Mathematics and Statistics, University of Western Australia, for their hospitality during Autumn 1998.

References.

Chang and Pollard (1997) Conditioning as disintegration, *Statistica Neerlandica* **51**, 287–317.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Sec. 9. Translated (1990) in *Statistical Science* **5**, 465–480.

Pearl, J. (1995). Causal Diagrams for Empirical Research, *Biometrika* **82**, 669–688.

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

Robins, J.M. (1987). Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect", *Computers and Mathematics with Applications* **14**, 923–945.

Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, pp. 113-159 in: *Health Service Research Methodology: A Focus on AIDS*, Sechrest, L., Freeman, H., and Mulley, A. (eds), NCHSR, U.S. Public Health Service.

Robins, J.M. (1997). Causal inference from complex longitudinal data, pp. 69–117 in: *Latent Variable Modeling and Applications to Causality*, M. Berkane (ed.), Lecture Notes in Statistics **120**, Springer.

Robins, J.M. and Wasserman L. (1997). Estimation of Effects of Sequential Treatments by Reparameterizing Directed Acyclic Graphs. pp. 409–420 in: *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, Providence Rhode Island, August 1–3, 1997, D. Geiger and P. Shenoy (eds.), Morgan Kaufmann, San Francisco.

Rubin, D.B. (1974). Estimating causal effects of treatment in randomized and non-randomized studies, *J. Educational Psychology* **66**, 688–701.

Rubin, D.B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization, *Ann. Statist.* **6**, 34–58.

Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*, New York: Springer Verlag.