The Annals of Statistics
2001, Vol. 29, No. 6, 1785–1811

CAUSAL INFERENCE FOR COMPLEX LONGITUDINAL DATA: THE CONTINUOUS CASE

BY RICHARD D. GILL AND JAMES M. ROBINS[1]

University Utrecht and Eurandom and Harvard School of Public Health

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. We also show that our assumptions concerning counterfactual variables place no restriction on the joint distribution of the observed variables: thus in a precise sense, these assumptions are "for free," or if you prefer, harmless.

1. Introduction.

1.1. *Preface.* Can we determine causality from correlations found in complex longitudinal data? According to Robins (1986, 1987, 1989, 1997) this is possible under certain assumptions linking the variables observed in the real world to variables expressing what would have happened, if.... Such variables are called counterfactuals. The very notion of counterfactuals has a checkered history in the philosophy of science and foundations of statistics.

In this paper we consider two fundamental issues concerning Robins' theory. First, do his assumed relations (between observed and unobserved—factual and counterfactual—random variables) place restrictions on the distribution of the observed variables. If the answer is *yes*, adopting his approach means making restrictive implicit assumptions, not very desirable. If, however, the answer is *no*, his approach is neutral. One can freely use it in modelling and estimation, exploring the consequences (for the unobserved variables) of the model. This follows the highly successful tradition in all sciences of making thought experiments. In what philosophical sense counterfactuals actually exist seems to us less relevant. However, it is important to know if a certain thought experiment is a priori ruled out by existing data.

Secondly, can this theory be extended from discrete to continuous data? If *no* there is again a fundamental barrier to flexible use of these models. Well, it turns out that there is a fundamental difficulty in this extension since the theory is built on conditioning, and conditional probability distributions are not uniquely defined for continuous variables. Simply rewriting the assumptions

Received March 1999; revised May 2001.
[1]Supported in part by NIH Grant AI32475.
AMS 2000 *subject classifications*. Primary 62P10; secondary 62M99.
Key words and phrases. Causality, counterfactuals, longitudinal data, observational studies.

for continuous variables and copying the proofs of the main results of the theory produces nonsense, since the results depend dramatically on arbitrary choices of versions of conditional distributions. One arrives at true mathematical theorems that have no practical content whatsoever.

We show that the approach can be saved under certain continuity assumptions, which enable a canonical choice of conditional distributions, having empirical content. The main theorems are shown to hold under quite natural adaptations of the main assumptions. Applied statisticians might counter that all data is really discrete, hence these difficulties are imaginary. However, in our opinion if a theory fails to extend from discrete to continuous, it is going to be useless in practice when one has many discrete variables with many levels. Grouping might be one solution, but it only makes sense under some kind of continuity.

The paper will be concerned throughout with these fundamental (probabilistic) aspects of Robins' theory; statistical issues (what to do with a finite sample of data, assuming these models) are not considered. However we start in Section 1.2 with a verbal description of a real world (like) example to help motivate the reader. In Section 2 we will then summarize the discrete time theory and specify the precise mathematical problems we are going to study. Section 3 collects a number of useful facts about conditional distributions. Next, in Section 4 we establish our main theorem generalizing the g-computation formula from discrete to continuous variables. In Section 5 we prove a number of characterizations of the null-hypothesis of no treatment effect. The motivation for proving these results is statistical, and connected to an approach to statistical modelling based on Structural Nested Distribution Models, briefly discussed at the beginning of the section. Then, in Section 6 we turn to the problem of showing that the counterfactuals of the theory are 'free', in the sense of placing no restrictions on the distribution of the observed variables. Finally, Section 7 gives an alternative approach to the continuity problem, based on randomized treatment plans.

1.2. *An Example.* Consider a study of the treatment of AIDS patients, in which treatment consists of medication at two separate time points t_1 and t_2, the second subtreatment depending on the state of the patients measured some time after the first time point. At a later time point still, t_3, the final health status of the patients is observed. The question is whether the treatment as a whole (i.e., the two subtreatments together) has an effect on the final outcome. To keep matters simple, we will suppose that there is no initial covariate information on each patient.

To be specific, at time t_1, each patient is given some amount (measured in micrograms) of the drug AZT (an inhibitor of the AIDS virus). Denote this amount by the random variable A_1. The amount varies from patient to patient (either deliberately, because the study is a randomized trial, or simply because the study is an observational multicenter study and different doctors have different habits). In the subsequent weeks patients might develop the illness PCP (pneumocystis pneumonia); severity is measured by the partial pressure

of oxygen in the arterial blood in millimeters of mercury, a continuous variable called PaO_2. We denote it by L_2. Depending on the outcome, at time t_2 patients will be treated with the drug aerosolized pentamidine (AP). Again, varying amounts are used (measured again in micrograms). We denote this by the random variable A_2. Finally, at some later time still, t_3, the treatment so far is evaluated by observing some variable Y. For this argument, let us suppose it is simply the binary variable "alive or dead". Thus in time order we observe the variables A_1, L_2, A_2, Y, where A_1 and A_2 together constitute the treatment, while L_2 and Y are responses: the first an intermediate response, the second the final response of interest. (More realistically, there would also be an initial variable L_1 representing the patient's state before the first treatment, but that is supposed absent or irrelevant here.)

Now let us suppose we have no statistical problems, but actually know the joint probability distribution of the four variables (A_1, L_2, A_2, Y), which we have identified with AZT, PaO_2, AP, Survival. The first three are continuous, the last is binary. How can we decide whether treatment (AZT and AP) has an effect on final outcome Survival, and how can we measure this effect? Standard approaches fail, since the intermediate variable PaO_2 is both influenced by treatment, so should be marginalized, and a cause of treatment (a covariate), so should be conditioned on. In our approach, we decide this by introducing a random variable Y^g which represents what the outcome would have been, had the patient been treated according to a given treatment plan g. What is a treatment plan in the context of this example? Well, it is a specification of a fixed amount of treatment by AZT at the initial timepoint t_1, and then at time t_2 (after PaO_2 has been measured) by an amount of treatment by AP depending on (and only depending on) the observed level of PaO_2. We denote these two components by g_1 and g_2, where g_1 specifies a fixed amount a_1 of treatment by AZT, and g_2 is a function assigning an amount of treatment $a_2 = g_2(l_2)$ by AP, for each possible value l_2 of PaO_2. (By the way, since g_1 and g_2 are two parts of one treatment plan, the second component already "knows" the AZT treatment and hence depends on l_2 only, not also on a_1.) Does treatment have an effect? By our definition, yes, if and only if the distribution of Y^g is not independent of the treatment plan g.

In the theory it is shown how, at least for discrete variables, and under certain assumptions, the law of Y^g can be computed from the joint law of the data, for any given g. The formula (called the g-computation formula) for this probability law is the theoretical starting point of further characterizations of no treatment effect and for the specification of statistical models for the effect of treatment which lend themselves well to statistical analysis; thus this formula is the keystone to a versatile and powerful statistical methodology.

We state these assumptions verbally and after that give a verbal description of the conclusion (the formula) which follows from the assumptions.

The first main assumption is called the consistency assumption and states that whenever, by coincidence, patients are factually treated precisely as they would have been treated had plan g been in operation, then their actual and their counterfactual final outcomes are identical. In our example, this means

that for each ω for which $A_1(\omega) = g_1 = a_1$ and $A_2(\omega) = g_2(L_2(\omega))$, it should hold that $Y^g(\omega) = Y(\omega)$. The consistency assumption is the generalization to time-varying treatments of both Cox's (1958) "no interaction between units" assumption that one subject's counterfactual outcomes do not depend on the treatment received by any other subject and Rubin's (1978) assumption that there is only one version of the treatment. Rubin (1978) refers to the combination of these two assumptions as the stable treatment unit value assumption (SUTVA). For continuously distributed treatments and intermediate outcomes, the consistency assumption concerns (for each g separately) an event of probability zero. Thus it is not a distributional assumption at all. We could make it true or false as we liked, by modifying our probability space on an event of probability 0. By the way, in the original discrete-variables theory one could have weakened this assumption to the assumption that the conditional law of Y^g given $A_1 = a_1$ and $A_2 = g_2(L_2)$ is equal to the conditional law of Y under the same conditioning. The assumption is now an assumption concerning a conditional law given an event of probability 0, hence again is meaningless.

The second main assumption is called the sequential randomization assumption, and also known as the assumption of *no unmeasured confounders*. It splits into two subassumptions, one for each of the two treatment times t_1, t_2. It states first that A_1 is independent of Y^g and second that A_2 is conditionally independent of Y^g given A_1 and L_1. This means that to begin with, the amount of treatment by AZT actually received does not in any way predict survival under any prespecified treatment plan g. This would be true in a randomized clinical trial by design. It may also be true in an observational study. But if doctors' treatment of patients depended on their unrecorded state of health and hence future prospects beyond what has been registered in our database, we are in bad trouble. In our example there is no information about the patient's state before the first treatment, and we are forced to assume that treatment did not depend on missing information. At the second time point, we must again assume that *given what is now known* (the initial treatment by AZT and health status as measured by PaO_2), the actual dose of AP does not depend on the patient's counterfactual outcome under plan g.

We see that these assumptions are also formulated in terms of conditional distributions given events of probability zero. Thus in the continuous case they are meaningless.

Now the aim is to arrive, under these assumptions, at the following theorem: the law of Y^g is the same as the law obtained in the following simulation experiment: fix $a_1 = g_1$, next draw a value l_2 from the conditional law of L_2 given $A_1 = a_1$, set $a_2 = g_2(l_2)$, and then finally draw a value y^g from the conditional law of Y given $(A_1, L_2, A_2) = (a_1, l_2, a_2)$. The formal version of this verbal solution is called the g-computation formula. As we stated it is the keystone of a whole statistical theory, but it depends in the continuous case on a collection of arbitrary conditional distributions given some zero probability events.

Let us briefly indicate the solution we have chosen, under a further simplification of the example. Suppose now that the only variables involved are

A_2, which we abbreviate to A, and Y. Suppose A is uniformly distributed on [0, 1] while the conditional law of Y given $A = a$ is degenerate at 0 for $a < \frac{1}{2}$, and degenerate at 1 for $a > \frac{1}{2}$. Of course the law of Y given $A = a$ can be arbitrarily altered on a set of measure zero, and in particular we need not give it at all for $a = \frac{1}{2}$. Suppose we are interested in the distribution of the outcome under the treatment plan $g = \frac{1}{2}$. Now random variables Y and A with this joint distribution live on the following probability space, together with a pair of rather different counterfactual outcomes Y^g and $Y^{g'}$, which however *both* satisfy the consistency and randomization assumptions. We take Ω to be $[0, 1] \cup \frac{1}{2}'$: the unit interval together with an extra point $\frac{1}{2}'$. Think of this set as being ordered as $[0, \frac{1}{2})$, $\frac{1}{2}$, $\frac{1}{2}'$, $(\frac{1}{2}, 1]$. We put the uniform distribution on this set and we define $A(\omega) = \omega$ if $\omega \in [0, 1]$, $A(\frac{1}{2}') = \frac{1}{2}$; $Y(\omega) = 0$ if $\omega \leq \frac{1}{2}$, $Y(\omega) = 1$ if $\omega \geq \frac{1}{2}'$; $Y^g(\omega) = 0$ for all ω except $\omega = \frac{1}{2}'$, where $Y^g = 1$; $Y^{g'}(\omega) = 1$ for all ω except $\omega = \frac{1}{2}$, where $Y^{g'} = 0$.

In this crazy example, both Y^g and $Y^{g'}$ are independent of A; thus the randomization assumption is satisfied. On the set $A = \frac{1}{2}$, both are identically equal to Y so the consistency assumption is satisfied. Finally, the law of Y given $A = \frac{1}{2}$ can be taken to be anything, so it can be chosen to equal either the law of Y^g or that of $Y^{g'}$, so we can make the the g-computation formula (as it was verbally described above) correct. But obviously, the question what the outcome would have been, under treatment $\frac{1}{2}$ just should not be posed in this case. Equally obviously, it does seem reasonable to ask what the outcome would have been had the treatment been any other fixed value. Note that the law of Y given $A = a$ is continuous in a except at $a = \frac{1}{2}$.

So our approach will be to assume the existence of continuous versions of relevant conditional distributions. We will formulate the consistency and randomization conditions in terms of conditional distributions, and show that the g-computation formula is then valid. Moreover, under a natural identifiability condition, the formula defines a functional of the joint law of the data. We will go on to show that key theorems characterizing the null hypothesis of no treatment effect also continue to hold in the new set-up, thus providing a sound basis for a continuous variables theory of causal inference for complex longitudinal data. We make some further introductory remarks on this topic at the start of the relevant section (Section 5).

The above remarks pertained to the second aim of the paper: to extend the theory from discrete to continuous. The other aim is to to show that the assumptions linking hypothetical variables Y^g to the observed variables (A_1, L_2, A_2, Y) are "free" in the sense that they impose no restrictions on the joint distribution of the observed data. Mathematically speaking, we want to show that given a probability space with variables (A_1, L_2, A_2, Y) defined on it, we can (possibly after augmenting the probability space with some independent uniform variables in order to have some extra source of randomness), also define a collection of variables Y^g for every treatment plan g, satisfying the consistency and randomization conditions. Thus the counterfactual

approach is, at least as a thought experiment, always permissible. If, moreover, the assumptions in the approach are plausible from a subject matter point of view, then the approach will provide a sound basis for predictions of what would happen if various treatment plans were introduced on new patients from the same population as before.

We do this (in Section 6) in a topsy-turvy way, by first building a counterfactual universe and then afterwards constructing the real world from some portions of it. The construction is clearly nonunique, thus not only are the assumptions free, but they leave many options open.

2. 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 fixed, that is, nonrandom, 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 kth visit to the clinic. Robins' approach generalizes the time-independent point-treatment counterfactual approach of Neyman (1923) and Rubin (1974, 1978) to the setting of longitudinal studies with time-varying treatments and covariates. Robins (1997) discusses the relationship between his theory and causal theories based on directed acyclic graphs and nonparametric structural equation models due to Pearl (1995) and Spirtes, Glymour and Scheines (1993).

We assume the study 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.

We shall consider both randomized clinical trials and observational studies. In an observational study the treatment decision at the kth visit is not determined by a specified protocol but is the result of the doctor's and patient's personal decisions at that moment. The treatment A_k given at the kth visit will vary with patient, physician, and calendar time 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.

Robins (1986, 1987, 1989, 1997) formally has proved results for this theory only for the case in which the covariates and treatments take values in discrete spaces. Our aim here is to extend these results 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. One of our motivations 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. Lok (2001) develops a counting process framework, within which she is able to formalize parts of the theory and prove many of the key results. It is an open problem to complete this project with a continuous time version of the g-computation formula and the theorems centered around it, which we study here.

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 kth defined on sequences of the first k covariate values, where $a_k = g_k(\bar{l}_k)$ is the treatment to be administered at the kth visit given covariate values $\bar{l}_k = (l_1, \ldots, l_k)$ up till then. Following the notational conventions already introduced, we define $\bar{g}_k(\bar{l}_k) = (g_1(l_1), g_2(l_1, l_2), \ldots, g_k(l_1, \ldots, l_k))$ and $\bar{g}(\bar{l}) = \bar{g}_K(\bar{l}_K)$. However for brevity we often abbreviate $\bar{g}_k$ or $\bar{g}$ simply to g when the context makes clear which function is meant, as in $\bar{a}_k = g(\bar{l}_k)$ or $\bar{a} = g(\bar{l})$.

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 as follows.

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

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

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

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 randomization assumption A2 states that the kth 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 identifiability 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,

$$\begin{aligned}\Pr(Y^g \in \cdot) = &\int_{[a_1=g_1(l_1)]} \cdots \int_{[a_K=g_K(\bar{l}_K)]} \Pr(Y \in \cdot \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K) \\ &\times \prod_{k=1}^{K} \Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}).\end{aligned} \tag{1}$$

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 asterisk 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 $\overline{L} = \bar{l}^*, \overline{A} = \bar{a}^*$.

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

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

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

It is important to note that we do not need Assumption A2† in proving (1) and that (1) can be valid without its obvious probabilistic interpretation B1 being correct. Some researchers take the probabilistic interpretation of (1) as being so natural that for them it is a primitive assumption. However, Robins [(1997), Section 7, pages 81–83] has given substantive examples where A2 would hold and (1) is therefore valid, but neither A2† nor B1 hold for certain choices of g. Informally, this will occur when, in the parlance of epidemiologists, there is an unmeasured variable U_0 that is a confounder for the effect of treatment on a component V_m of L_m but does not confound the effect of treatment on the outcome Y as for example when (a) U_0 and A_{k+1} are conditionally dependent given $(\overline{L}_k, \overline{A}_k)$ for some k, (b) U_0 is a cause of V_m for each m, (c) neither U_0 nor V_m is a cause of the outcome Y, (d) $g(\bar{l}_m)$ is the same for all $\bar{v}_m$ and (e) Assumption A2† would hold were U_0 included as an element of L_0. In conclusion, we believe that (1) needs to be motivated on subject matter grounds, and that conditions A1 to A3 are both meaningful and as weak as possible for this job.

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 Kth 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, that is, 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 $(\bar{l}_k, \bar{a}_k)$ which are of the form $(\bar{l}_k, \bar{g}_k(\bar{l}_k))$ 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 Section 4, after some technical groundwork in Section 3, we will state the 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 as easy an exercise though it still follows the same line as given above.

We go on in Section 5 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.

In Section 6 we turn to this question: given factual variables $(\overline{L}, \overline{A}, Y)$ can one construct a variable Y^g satisfying A1 and 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 hand 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 (and it does), that is fine.

We emphasize that the correctness of (1) and the uniqueness of (the right-hand side) of (1) are two different issues. We saw at the end of Section 1.2 a small artificial example where there are two different counterfactual variables Y^g and $Y^{g'}$, with different marginal distributions, both satisfying A1 and 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. 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.

In Section 7 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 nondistributional assumptions on the underlying probability space. This approach raises some interesting open problems.

3. Facts on conditioning.

Conditional distributions. We assume without further mention from now on that all variables take values in Polish spaces (that is, 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

$$\Pr(X \in dx, Y \in dy) = \Pr(X \in dx \mid Y = y)\Pr(Y \in dy). \tag{2}$$

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). See Pollard (2001) for a modern treatment of conditional distributions.

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:

$$\int \Pr(X \in \cdot \mid Y = y, Z = z)\Pr(Y \in dy \mid Z = z) = \Pr(X \in \cdot \mid Z = z).$$

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 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, that is, it has probability 1 (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, Supp(X), is closed and has probability 1.

PROOF. Any point not in the support is the center of an open ball of probability 0. 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 0, hence it has probability 0. □

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

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

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

$$\begin{aligned}\mathrm{E}(f(X) \mid Y \in B(y_0, \delta)) = \int_{B(y_0, \delta) \cap \text{Supp}(Y)} & \mathrm{E}(f(X) \mid Y = y) \\ & \times \Pr(Y \in dy \mid Y \in B(y_0, \delta)),\end{aligned}$$

where $\mathrm{E}(f(X) \mid Y = y)$ 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 $\text{law}(Y \mid Y \in B(y_0, \delta)) \to \delta_{y_0}$ as $\delta \downarrow 0$, the right hand side converges to $\mathrm{E}(f(X) \mid Y = y_0)$ as $\delta \downarrow 0$. □

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

PROOF. For $y \in \text{Supp}(Y)$ and $x \in \text{Supp}(X \mid Y = y)$ 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 $\Pr(Y \in B(y, \varepsilon)) > 0$. However,

$$\begin{aligned}\Pr\big((X, Y) \in B(x, \delta) \times B(y, \delta)\big) \geq & \Pr(Y \in B(y, \varepsilon)) \\ & \times \Pr(X \in B(x, \delta) \mid Y \in B(y, \varepsilon))\end{aligned}$$

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

One might expect that the union over $y \in \text{Supp}(Y)$ of the sets $\text{Supp}(X \mid Y = y) \times \{y\}$ is precisely equal to $\text{Supp}(X, Y)$ but this is not necessarily the case. The resulting set can be strictly contained in $\text{Supp}(X, Y)$ though it is *a* support of (X, Y) in the sense of having probability 1. Its closure equals $\text{Supp}(X, Y)$.

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 $(y, z) \in \text{Supp}(Y, Z)$. Suppose $\Pr(Y \in$

$\cdot \mid Z = z)$ is continuous in $z \in \mathrm{Supp}(Z)$. Then

$$\Pr(X \in \cdot \mid Z = z) = \int_y \Pr(X \in \cdot \mid Y = y, Z = z) \Pr(Y \in dy \mid Z = z)$$

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

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

as $z \to z_0$, $z \in \mathrm{Supp}(Z)$. Suppose without loss of generality that $|f|$ is bounded by 1. The function $g(y, z) = \mathrm{E}(f(X) \mid Y = y, Z = z)$, 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 $\mathrm{E}(f(X) \mid Y = y, Z = z)$ 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

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

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 $\mathrm{E}(f(X) \mid Y = y, Z = z)$ 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, $\int \mathrm{E}(f(X) \mid Y = y, Z = z) \Pr(Y \in dy \mid Z = z)$ is within 2ε of $\int_K \mathrm{E}(f(X) \mid Y = y, Z = z_0) \Pr(Y \in dy \mid Z = z)$. Again for z close enough to z_0, this is within 3ε of $\int \mathrm{E}(f(X) \mid Y = y, Z = z_0) \Pr(Y \in dy \mid Z = z)$. Since the integrand here is a fixed bounded continuous function of y, for $z \to z_0$ this converges to $\int \mathrm{E}(f(X) \mid Y = y, Z = z_0) \Pr(Y \in dy \mid Z = z_0)$. Thus for z close enough to z_0, $\int \mathrm{E}(f(X) \mid Y = y, Z = z) \Pr(Y \in dy \mid Z = z)$ is within 4ε of $\int \mathrm{E}(f(X) \mid Y = y, Z = z_0) \Pr(Y \in dy \mid Z = z_0)$. □

4. 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 generalization of condition A3.

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

As with the original version of A3, the condition calls the result of a plan g identifiable 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.

ASSUMPTION C (Continuity). The distributions $\text{law}(Y \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K)$ can be chosen continuous in $(\bar{l}_K, \bar{a}_K)$, and $\text{law}(L_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1})$ in $(\bar{l}_{k-1}, \bar{a}_{k-1})$, 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 side of (1) has the probabilistic interpretation that first a value l_1 is generated according to $\text{law}(L_1)$, then a_1 is specified by $a_1 = g_1(l_1)$, then a value l_2 is generated from $\text{law}(L_2 \mid L_1 = l_1, A_1 = a_1)$, and so on. Suppose that at the end of the kth step we have obtained $(\bar{l}_k, \bar{a}_k) \in \text{Supp}(\overline{L}_k, \overline{A}_k)$. Then l_{k+1} will with probability 1 be generated, according to a uniquely determined probability distribution, in $\text{Supp}(L_{k+1} \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k)$, thus $(\bar{l}_{k+1}, \bar{a}_k) \in \text{Supp}(\overline{L}_{k+1}, \overline{A}_k)$ by Fact 3. By condition A3*, this leads to $(\bar{l}_{k+1}, \bar{a}_{k+1}) \in \text{Supp}(\overline{L}_{k+1}, \overline{A}_{k+1})$. By induction, with probability 1 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 $b(g; \text{law}(\overline{L}, \overline{A}, Y))$ on the right-hand side of (1) is well defined. We next want to investigate when it equals $\text{law}(Y^g)$. 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.

ASSUMPTION Cg (Continuity for counterfactuals). The distributions $\text{law}(Y^g \mid \overline{L}_{k+1}, \overline{A}_k)$ and $\text{law}(Y^g \mid \overline{L}_k, \overline{A}_k)$ 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 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:

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

ASSUMPTION A2* (Randomization). Suppose law($Y^g \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k$) does not depend on a_k for $\bar{a}_k, \bar{l}_k \in \text{Supp}(\overline{L}_k, \overline{A}_k)$ 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 $(\bar{l}_K, \bar{a}_K) \in \text{Supp}(\overline{L}_K, \overline{A}_K)$ and $g(\bar{l}_K) = \bar{a}_K$, where both conditional distributions are uniquely determined by continuity. Now let $(\bar{l}_{k-1}, \bar{a}_{k-1}) \in \text{Supp}(\overline{L}_{k-1}, \overline{A}_{k-1})$ and satisfying $g(\bar{l}_{k-1}) = \bar{a}_{k-1}$ be fixed. Consider

$$\begin{aligned} &\int_{l_k \in \text{Supp}(L_k \mid \overline{L}_{k-1} = \bar{l}_{k-1},\, \overline{A}_{k-1} = \bar{a}_{k-1}) \atop [a_k = g_k(\bar{l}_k)]} \Pr(Y^g \in \cdot \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k) \\ &\qquad \times \Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}). \end{aligned} \tag{4}$$

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

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

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 law($L_k \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1}$) is continuous in $(\bar{l}_{k-1}, \bar{a}_{k-1})$, but also the following.

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

Then for Cg it suffices to assume that law($Y^g \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K$) is continuous in $(\bar{l}_K, \bar{a}_K)$ 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 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, $\Pr(Y = Y^g \mid \overline{L} = \bar{l}, \overline{A} = \bar{a}) = 1$ for $\bar{a} = g(\bar{l})$. 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 $\mathscr{G}$ of all plans g. So far we were not able to settle this question. It arises again when we consider the alternative approach based on randomized plans in Section 6.

5. Characterizing the null-hypothesis. The statistician's first interest in applications of this theory, working with an i.i.d. sample from the distribution of $(\overline{L}, \overline{A}, Y)$, would probably be to test the null hypothesis of no-treatment effect, and secondly to model the effect of treatment and estimate the model parameters involved in the effect of treatment. For example, were the null hypothesis rejected, one might wish to test the p-latent period hypothesis that only treatments received more than p time periods prior to the study end at K have an effect on Y. Unfortunately the g-computational formula as it stands does not lend itself very well to these aims. Even just to estimate $b(g)$ for a single plan g would appear to involve estimation of a large number of high-dimensional nonnonparametric regessions followed by a Monte Carlo experiment using the estimates. How to test equality of the $b(g)$ over all possible g seems even less feasible.

Typically one introduces parametric or semiparametric models to alleviate problems due to the curse of dimensionality. Thus one might consider a parametric specification of each conditional law involved in the g-computational formula. This might make estimation closer to feasible; however, it does not aid in the testing problem, since the null hypothesis is now specified by a very complex functional of all parameters. Since the parametric models for the ingredients of $b(g)$ will usually be no better than rough approximations, the null hypothesis will for large samples be rejected simply through massive specification error.

In order to solve these problems, Robins (1986, 1987, 1989, 1997) derived alternative characterizations of the "g"-null hypothesis H_0 that the right-hand side of (1) is the same for all identifiable treatment plans g. The alternative characterizations also provide a starting point for modelling and estimation using so-called structural nested distribution models. It is therefore important to see whether these results too can be carried over to the continuous case. By an identifiable plan g we now mean a plan satisfying the identifiability assumption A3*. The characterization 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 identifiable plan g is the same, and hence treatment indeed has no effect on outcome.

Theorems 3 and 4 below give an initial simplification of the testing problem. Theorem 5 goes further in showing 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 (nonsaturated), then typically no combination of parameter values could even reproduce the null-hypothesis [Robins (1997), Robins and Wasserman (1997)]. Theorems 4 and 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 a certain 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 mutually 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.

We call a treatment plan static if it does not depend in any way on the covariate values $\bar{l}$; 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 $g \equiv \bar{0} = (0, \ldots, 0)$ 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 $\bar{0}$, are not identifiable. (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 identifiable 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 identifiable. 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 $(\bar{l}_k, \bar{a}_{k-1})$, introduce the quantity

$$b(g;\bar{l}_k,\bar{a}_{k-1}) = \int_{l_{k+1}} \cdots \int_{l_K} \Pr\big(Y \in \cdot \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K\big) \times \prod_{k'=k+1}^{K} \Pr(L_{k'} \in dl_{k'} \mid \overline{L}_{k'-1} = \bar{l}_{k'-1}, \overline{A}_{k'-1} = \bar{a}_{k'-1}\big), \tag{6}$$

where $a_k, \ldots, a_K$ on the right-hand side are taken equal to $g_k(\bar{l}_k), \ldots, g_K(\bar{l}_K)$. Similarly to Theorem 1, this is a well-defined functional of the joint law of the factual variables when $(\bar{l}_k, \bar{a}_{k-1})$ lies in the support of $(\overline{L}_k, \overline{A}_{k-1})$, when $g(\bar{l}_{k-1}) = \bar{a}_{k-1}$ and when g is identifiable, 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 kth, so it is well-defined more generally than this. Let us say that a plan g is k-identifiable relatively to a given $(\bar{l}_k, \bar{a}_{k-1})$ if for all $m \geq k$, any $(\bar{l}_m, \bar{a}_{m-1}) \in \mathrm{Supp}(\overline{L}_m, \overline{A}_{m-1})$ 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 $(\bar{l}_m, \bar{a}_m) \in \mathrm{Supp}(\overline{L}_m, \overline{A}_m)$ where of course $g_m(\bar{l}_m) = a_m$.

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

The theorems we want to prove are the following:

THEOREM 3. *Assume condition* C *and the null hypothesis* H_0: *equality of* $b(g)$ *for all identifiable plans* g. *Then 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*-identifiable plan* g.

THEOREM 4. *Assume condition* C. *Suppose an admissible baseline plan* g^0 *exists. Then if for all* $(\bar{l}_k, \bar{a}_k)$ *in the support of* $(\overline{L}_k, \overline{A}_k)$ *the expression* $b(g^{k+1:0};\bar{l}_k, \bar{a}_{k-1})$ *does not depend on* $a_k = g_k(\bar{l}_k)$, 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 $g_k(\bar{l}_k)$. 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-identifiable 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 $g_0 \equiv \bar{0}$ is an identifiable 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. *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 blip-down functions* $\gamma_k = \gamma_k(y;\bar{l}_k, \bar{a}_k)$ *can be found*

satisfying the following: *if 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. *Furthermore the right-hand side of* (1) *is given by*

$$\int \cdots \int_{\{h(y^0; \bar{l}_K, \bar{a}_K) \in \cdot\}} \prod_k \Pr(L_k \in dl_k | Y^0 = y^0, \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}) \Pr(Y^0 \in dy^0),$$

where the "blip all the way up" function $h = \gamma_K^{-1} \circ \gamma_{K-1}^{-1} \circ \cdots \circ \gamma_1^{-1}$, *and the "blip up" function* γ_k^{-1} *is the inverse of the "blip down" function* γ_k *with respect to its first argument*.

Consider a setting in which 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.

Furthermore, suppose the suppositions of Theorem 2 hold and $g^0 \equiv \bar{0}$. Then $\gamma_k(y; \bar{l}_k, \bar{a}_k)$ represents (on a quantile–quantile map scale) the effect on the subset of subjects with history $(\bar{l}_k, \bar{a}_k)$ of one final blip of treatment of magnitude a_k on subjects with observed history $\overline{L}_k = \bar{l}_k$ and $\overline{A}_k = \bar{a}_k$. Further, H_0 holds if and only if $\gamma_k(y; \overline{L}_k, \overline{A}_k) = y$ almost surely for all k and y. Similarly, the p-latent period hypothesis holds if and only if $\gamma_k(y; \overline{L}_k, \overline{A}_k) = y$ almost surely for all $k \geq K - p$, y. In practice, to test both H_0 and the p-latent period hypotheses, we could specify a parametric model $\gamma_k^*(y; \bar{l}_k, \bar{a}_k, \psi)$ for $\gamma_k(y; \bar{l}_k, \bar{a}_k)$ where $\gamma_k^*(y; \bar{l}_k, \bar{a}_k, \psi)$ is a known function satisfying $\gamma_k^*(y; \bar{l}_k, \bar{a}_k, \psi) = y$ whenever $a_k = 0$ and $\psi = (\psi_1, \psi_2)$ is a finite-dimensional unknown parameter with true value ψ^* satisfying $\psi = 0$ if and only if $\gamma_k^*(y; \bar{l}_k, \bar{a}_k, \psi) = y$ for all $y, \bar{l}_k, \bar{a}_k$, $k = 1, \ldots, K$ and $\psi_2 = 0$ if and only if $\gamma_k^*(y; \bar{l}_k, \bar{a}_k, \psi) = y$ for all $y, \bar{l}_k, \bar{a}_k$, $k = K - p, K - p + 1, \ldots, K$. Then $\psi = 0$ if and only if H_0 holds and $\psi_2 = 0$ if and only if the p-latent period hypothesis holds.

Now suppose for concreteness that A_k is a Bernoulli random variable and we have a correctly specified linear logistic model $\Pr(A_k = 1 \mid \overline{L}_k, \overline{A}_{k-1}) = \operatorname{expit}(\alpha^\top W_k)$, $k = 1, \ldots, K$, where α is an unknown parameter vector, W_k is a known function of $\overline{L}_k, \overline{A}_{k-1}$, and $\operatorname{expit}(x) = e^x/(1 + e^x)$. To estimate ψ^* we let $Y^0(\psi)$ be Y^0 but with $\gamma_k^*(y, \bar{l}_k, \bar{a}_k, \psi)$ substituted for $\gamma_k(y; \bar{l}_k, \bar{a}_k)$. Since, under our assumptions, $Y^0(\psi^*) \perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$ for all k, we estimate ψ^* as the value $\hat{\psi}$ of ψ for which the MLE of θ in the expanded model $\Pr(A_k = 1 \mid \overline{L}_k, \overline{A}_{k-1}, Y^0(\hat{\psi})) = \operatorname{expit}(\alpha^\top W_k + \theta Y^0(\hat{\psi}))$ is zero, where $\hat{\psi}$ is regarded as fixed when maximizing the logistic likelihood over (θ, α).

Finally we discuss the importance of using the formula given in Theorem 6 to compute the right-hand side of (1). Suppose $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2)$ with $\hat{\psi}_2 = 0$. By calculating $\Pr(Y^g \in \cdot)$ using the formula in Theorem 6 with $h(y_0; \bar{l}_K, \bar{a}_K, \hat{\psi})$ substituted for $h(y_0; \bar{l}_K, \bar{a}_K)$, our estimates of $\Pr(Y^{g^1} \in \cdot)$ and $\Pr(Y^{g^2} \in \cdot)$ will be equal whenever g^1 and g^2 agree through $K - p - 1$, that is, whenever $g_k^1 = g_k^2$, $k = 1, \ldots, K - p - 1$, regardless of how we estimate $\Pr(Y_0 \in dy_0)$

or $\Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}, Y_0 = y_0)$. Hence our estimate of $\Pr(Y^g \in \cdot)$ is guaranteed to be consistent with our hypothesized latent period of at least p time periods.

PROOF OF THEOREM 3. Suppose H_0 is true. Consider two plans g^1 and g^2. We want to prove equality of $b(g^i; \bar{l}_k^0, \bar{a}_{k-1}^0)$ 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 $a_1^0, \ldots, a_{k-1}^0$ 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 $\bar{l}_m$ for which $\bar{l}_k$ is in an epsilon ball about $\bar{l}_k^0$, 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 $\bar{l}_k$ in the epsilon ball about $\bar{l}_k^0$ which is precisely where g^1 and g^3 differ; denote this set $B(\bar{l}_k^0, \varepsilon)$. Deleting the integrals over the complement of this set we obtain the equality, for $i = 1, 2$, of the two quantities

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

Now by our continuity assumptions and repeated use of Fact 4, $b(g^i; \bar{l}_k, \bar{a}_{k-1}^0)$ is a continuous function of $\bar{l}_k$. Divide (7) by the normalizing quantity

$$\int_{\bar{l}_k \in B(\bar{l}_k^0, \varepsilon)} \prod_1^k \Pr(L_j \in dl_j \mid \overline{L}_{j-1} = \bar{l}_{j-1}, \overline{A}_{j-1} = \bar{a}_{j-1}^0),$$

the same for both $i = 1, 2$. Now the equality expresses the equality of the expectations of $b(g_i; \overline{L}_k^\varepsilon, \bar{a}_{k-1}^0)$ for $i = 1, 2$ where $\overline{L}_k^\varepsilon$ lies with probability one in $B(\bar{l}_k^0, \varepsilon)$. As $\varepsilon \to 0$, by continuity of $b(g_i; \cdot, \bar{a}_{k-1}^0)$, the expectations converge to $b(g_i; \bar{l}_k^0, \bar{a}_{k-1}^0)$. □

PROOF OF THEOREM 4. Let g be a given identifiable 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 kth treatment equal to the fixed amount a_k and all subsequent treatments follow the baseline plan. We show by downwards induction on k that $b(g; \bar{l}_k, \bar{a}_{k-1}) = b(g^{k:0}; \bar{l}_k, \bar{a}_{k-1})$ for all k. This statement for $k = 0$ is the required conclusion. To initialize the induction note that $b(g; \bar{l}_K, \bar{a}_{K-1}) = b(g^{K+1:0}; \bar{l}_K, \bar{a}_{K-1}) = b(g^{K:0}; \bar{l}_K, \bar{a}_{K-1})$, where the first equality is trivial and the second is the assumption of the theorem for $k = K$. Next, in general, write

$$\begin{aligned} b(g; \bar{l}_k, \bar{a}_{k-1}) &= \int_{l_{k+1}} b(g; \bar{l}_{k+1}, \bar{a}_k) \Pr(L_{k+1} \in dl_{k+1} \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k) \\ &= \int_{l_{k+1}} b(g^{k+1:0}; \bar{l}_{k+1}, \bar{a}_k) \Pr(L_{k+1} \in dl_{k+1} \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k) \end{aligned}$$

$$
\begin{aligned}
&\text{(by the induction hypothesis)}\\
&= b(g^{k+1:0}; \bar{l}_k, \bar{a}_{k-1})\\
&= b(g^{k:g_k(\bar{l}_k),0}; \bar{l}_k, \bar{a}_{k-1}) \qquad \text{(by inspection)}\\
&= b(g^{k:0}; \bar{l}_k, \bar{a}_{k-1}) \qquad \text{(by the assumption of the theorem)}
\end{aligned}
$$

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 identifiable 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 identifiable 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 4 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.

Note that the right-hand side of (1) is the probability $\Pr(Y \in \cdot)$ based on a new distribution in which the conditional distribution of Y given the past is unchanged, the conditional distribution of L_k given the past is unchanged, but the conditional distribution of A_k given $\overline{L}_k, \overline{A}_{k-1}$ is now degenerate according to g_k. Next note that, in the original distribution, $T = h(Y_0; \overline{L}_K, \overline{A}_K)$ which is one-to-one in its first argument, so we can rewrite the original probability distribution of $T, \overline{L}_K, \overline{A}_K$ as a transformation of that of $Y_0, \overline{L}_K, \overline{A}_K$. Express this latter as in the following simulation experiment: draw Y_0, draw L_k given the past including Y_0, then A_k given the past until K but note that the distribution of A_k given the past does not depend on Y_0 by our independence theorem. Now to get the new distribution needed to compute the right-hand side of (1) we again replace the law of $A_k \mid \overline{L}_k, \overline{A}_{k-1}$ by the degenerate law, and we have exactly the form in Theorem 6. □

6. Construction of counterfactuals. Suppose we start with a given law$(\overline{L}, \overline{A}, Y)$. Can we build on a possibly larger 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 in the discrete case (or the strengthened versions of these assumptions, in the general case)? The answer will be yes, in complete generality. 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.

We proceed to describe our construction in complete generality, and afterwards explain what it actually achieves, distinguishing between the discrete case in which we are interested in the original Assumptions A1, A2 and A3, and the general case, for which the assumptions need to be reformulated.

The construction works in the opposite direction to what one would expect: we construct a counterfactual world first on a completely new sample space, then build a copy of the factual world on top of it. Once we have constructed all variables together with the required properties, including the factuals with their given distribution, we can read off 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 $\bar{l}$; that is, it is just a single sequence of treatments a_k to be applied irrespective of the measured covariate values. Let $\mathscr{G}_0$ denote the collection of static plans; it can be identified with the collection of all $\bar{a}$.

First we build random variables $\overline{L}^{g_0}, Y^{g_0}$ for all $g_0 \in \mathscr{G}_0$. 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 $\overline{L} = \bar{l}, \overline{A} = \bar{a}$ 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 $\overline{L}_k = \bar{l}_k$, $\overline{A}_{k-1} = \bar{a}_{k-1}$ on the event $\overline{L}_k = \bar{l}_k$, $\overline{A}_{k-1} = \bar{a}_{k-1}$. 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.

In the discrete case, the consistency assumption A1 holds by construction. The randomization assumption A2 holds in the very strong form $(Y^g : g \in \mathscr{G}) \perp A_k \mid \overline{L}_k, \overline{A}_{k-1}$ where $\mathscr{G}$ 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. The identifiability condition A3 depends on which plan g is being considered, and is a condition on the joint law of the factual variables only, so is not of interest for the present purposes.

The collection of conditional distributions we used at the start of the construction is not uniquely defined in general. Even in the discrete case, it is not uniquely defined if not all values of $\overline{L}$, $\overline{A}$ have positive probability. Moreover, as we made clear in earlier sections, Assumptions A1 and A2 in their original versions are not distributional assumptions, that is, they cannot be checked by looking at the joint law of the factual variables and counterfactual variables. Whether or not they are true, depends on choices of conditional distributions and on other features of a specific underlying sample space. However under the continuity conditions C on the factual variables, the alternative assumptions A1* and A2* are distributional assumptions. One can check that under condition C, if we have chosen all conditional distributions in the construction subject to continuity on the supports of the conditioning variables, then the construction satisfies the stronger conditions Cg, A1* and A2*.

7. The *G*-computation formula for randomized 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, we 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 randomized treatment plan now denoted by G consists of a sequence of conditional laws $\Pr(A_k^G \in \cdot \mid \overline{L}_k^G = \bar{l}_k, \overline{A}_{k-1}^G = \bar{a}_{k-1})$. (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

$$\begin{aligned}\Pr(Y^G \in dy) = &\int_{l_1}\int_{a_1}\cdots\int_{l_K}\int_{a_K} \Pr(Y \in dy \mid \overline{L}_K = \bar{l}_K, \overline{A}_K = \bar{a}_K) \\ &\times \prod_{k=1}^{K} \Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}) \\ &\times \Pr(A_k^G \in da_k \mid \overline{L}_k^G = \bar{l}_k, \overline{A}_{k-1}^G = \bar{a}_{k-1}).\end{aligned} \tag{8}$$

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.

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

THEOREM 7. *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

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

The successive integrations with respect 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 $\bar{l}_K, \bar{a}_K$ which can actually be realized.

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

$a_k^G(\bar{l}_k, \bar{a}_{k-1}; u_k)$. Tracking through all possible values of all L_k^G, we see that the randomized plan G is exactly equivalent to choosing in advance, by a randomization depending only on $U_1, \ldots, U_K$, a non-randomized plan $g = g_{\vec{u}}$. A little thought shows that the right-hand side of (6) can be rewritten as $\int \cdots \int b(g_{\vec{u}}; \text{law}(\overline{L}, \overline{A}, Y)) du_1 \cdots du_K$. So if we make the additional consistency assumption $Y^G = Y^g$ on $G = g$, then (8) gives a *correct* expression for 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 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), Pollard (2001)] 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.

ASSUMPTION A0** (Sample space regularity). The underlying probability space $(\Omega, \mathscr{F}, \Pr)$ 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,

$$\int_{a_k} \Pr(\cdot \mid \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k) \Pr(A_k \in da_k \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1}) = \Pr(\cdot \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1})$$

and similarly

$$\int_{l_k} \Pr(\cdot \mid \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1}) \Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}) = \Pr(\cdot \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}).$$

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 $\mathscr{G}$ the set of all plans g, let $Y^{\mathscr{G}}$ 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 $\mathscr{F}$ generated by all Y^g.

Consider the following versions of A1 and A2.

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

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

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

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

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

$$\begin{aligned}(10)\qquad &\int_{u_1} \cdots \int_{u_K} \int_{l_1} \cdots \int_{l_K} \Pr(Y \in \cdot \mid \overline{L} = \bar{l}, \overline{A} = \bar{a}) \\ &\quad \times \prod_{k=1}^{K} \Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}) du_1 \cdots du_K,\end{aligned}$$

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 $g = g_{\bar{u}}$. 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 $\Pr(Y^g \in \cdot \mid \overline{L} = \bar{l}, \overline{A} = \bar{a})$, at least, for almost all $\bar{l}$, $\bar{a}$. The exceptions do not, however, change the value of the integral. Moreover we can do this irrespective of the value of $g = g_{\bar{u}}$. Now we may mix over the conditional law of L_K, reducing (10) to

$$\begin{aligned}&\int_{u_1} \cdots \int_{u_K} \int_{l_1} \cdots \int_{l_{K-1}} \Pr(Y^g \in \cdot \mid \overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}) \\ &\quad \times \prod_{k=1}^{K-1} \Pr(L_k \in dl_k \mid \overline{L}_{k-1} = \bar{l}_{k-1}, \overline{A}_{k-1} = \bar{a}_{k-1}) du_1 \cdots du_K,\end{aligned}$$

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

$$\int_{u_1} \cdots \int_{u_K} \Pr(Y^{g_{\bar{u}}} \in \cdot) du_1 \cdots du_K = \Pr(Y^G \in \cdot). \qquad \square$$

The above theory is not 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. Fortunately, we were able to show in Theorem 7 that the main output of the theory, formula (8), is uniquely defined from the joint law of the data.

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

REFERENCES

CHANG, J. and POLLARD, D. (1997). Conditioning as disintegration. *Statist. Neerlandica* **51** 287–317.

COX, D. (1958). *The design and planning of experiments*. Chapman and Hall, London.

LOK, J. (2001). Statistical modelling of causal effects in time. Ph.D. thesis, Free Univ., Amsterdam. Available at www.cs.vu.nl/~jjlok.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: essay on principles, Sec. 9. Translated and reprinted in *Statist. Sci*. **5** 465–480.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–688.

POLLARD, D. (2001). *User's Guide to Measure-theoretic Probability*. Cambridge Univ. Press.

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

ROBINS, J. (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." *Comput. Math. Appl.* **14** 923–945.

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

ROBINS, J. (1997). Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statist*. **120** 69–117. Springer, Berlin.

ROBINS, J. and WASSERMAN, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, (D. Geiger and P. Shenoy, eds.) 409–420. Morgan Kaufmann, San Francisco.

RUBIN, D. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *J. Educational Psychology* **66** 688–701.

RUBIN, D. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist*. **6** 34–58.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search*. Springer, New York.

MATHEMATICAL INSTITUTE
UNIVERSITY UTRECHT
BOX 80010
3508 TA UTRECHT
NETHERLANDS
E-MAIL: gill@math.uu.nl

DEPARTMENT OF EPIDEMIOLOGY
HARVARD SCHOOL OF PUBLIC HEALTH
677 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
E-MAIL: robins@hsph.harvard.edu