
Chapter 28

EFFECTS OF MULTIPLE INTERVENTIONS

JAMES ROBINS, MIGUEL HERNAN AND UWE SIEBERT

1. INTRODUCTION

The purpose of this chapter is (i) to describe some currently available analytical methods for using individual level epidemiological data to estimate the impact of multiple risk factor interventions on health and (ii) to carefully review the conditions under which these methods deliver unbiased estimates of impact. The chapter is organized as follows. In sections 2 and 3, we discuss estimation of effects of short-term, time-independent interventions. Specifically, we discuss estimating the effect of a single risk factor intervention on life expectancy or quality-adjusted life expectancy over a specified period of follow-up in a single population, when essentially ideal epidemiologic data are available. That is, we assume a random sample of the population is randomly assigned to different levels of the risk factor and followed prospectively for a fixed time period. Second, we consider the same study design, except now we are interested in the joint effect of an intervention on several risk factors. Third, we consider the problem of extrapolation of the results to longer periods of follow-up and to other populations for which no primary epidemiologic data are available. Sections 2 and 3 serve to indicate the possibilities and limitations of even ideal epidemiologic data for estimating the effects of multiple time-independent risk factor interventions. In sections 4 and 5 we turn to the main topic of this chapter: the estimation of the effect of multiple time-dependent interventions from observational data, possibly plagued by confounding, selection bias, measurement error, information bias and ill-defined interventions. In sections 6 to 8, we illustrate our methods by attempting to estimate the effects of various time-varying interventions on subjects entered in the Framingham Offspring cohort study. Finally, in section 9 we offer some conclusions.

2. TIME-INDEPENDENT INTERVENTIONS

2.1 AN IDEAL INTERVENTION

We suppose we have data from a randomized trial of the effect of a one-time short-term intervention (say, an anti-smoking intervention involving one week of intense study of literature on the health consequences of smoking) initiated at calendar time, say 1983. We suppose the subjects in the trial constitute a random sample of the population of a given country, say the United States of America, and these subjects are followed for twenty years. Let $S_0(t)$ and $S_1(t)$ be the survival curves of the exposed and unexposed respectively at t years from the beginning of the trial for $t \leq 20$ years. Thus $S_0(t)$ and $S_1(t)$ are both equal to one at time $t = 0$ and decrease as t increases. Suppose there is no sharing of information in the sense subjects who receive the anti-smoking intervention do not pass on (i.e. infect or contaminate) others in the trial or in the general population with their acquired knowledge. Then it is well known that the area between the survival curves is equal to the expected years of life saved over the first twenty years of the trial due to the intervention. If we are interested in quality-adjusted years of life lost, we use subject-year-specific health and interview data to associate with each year a subject lives a quality measure (taking values between 0 and 1 where 1 indicates optimal quality) and compare expected quality-adjusted years of life lost.

2.2 MULTIPLE IDEAL INTERVENTIONS

Turn now to estimating the effect of simultaneous interventions on k time-independent risk factors $A = (A_1, \dots, A_k)$, such as smoking, alcohol, blood pressure, and cholesterol, etc., where for the moment we assume all interventions were randomly assigned at the same moment in time and only once. Here A is the k -vector of all risk factors. Each risk factor A_m has some number $|A_m|$ of possible levels $a_{m1}, \dots, a_{m|A_m|}$. Then A has $|A| = |A_1| \times |A_2| \times \dots \times |A_k|$ possible joint levels. Let the set $\mathcal{A} = \{a\}$ of size $|A|$ denote the set of possible values a of the vector A . Let $S_a(t)$ be the survival curve for the random sample of subjects assigned to joint level a of the various risk factors. Let $S(t)$ denote the survival curve for a random subset of the population on which no interventions were made. Then the expected years of life gained by intervening and setting each subject's level of the k risk factors to a compared to no intervention is precisely the area between the survival curves $S_a(t)$ and $S(t)$. The optimal intervention a^* is the value of a for which the area under $S_a(t)$ is largest. Further, the loss in life expectancy under intervention a compared to the optimal is simply the difference in area under $S_{a^*}(t)$ and $S_a(t)$. In principle, we need make no assumptions concerning necessary and sufficient causes, multiplicative or additive interactions, or the fraction of deaths attributable to any particular cause to order the health benefits of various joint interventions on life expectancy from such ideal

epidemiologic data. All we need is a way to accurately estimate $S_a(t)$ for each joint intervention a in \mathcal{A} . Due to random sampling variability, in order to obtain reliable estimates of each of the $|A|$ curves $S_a(t)$ would require an inconceivably large randomized trial if $|A|$ was at all large, since a large number of subjects would need to be randomly assigned to each of the $|A|$ possible levels of A . In practice such large trials are infeasible. As a consequence, we can randomize subjects to only a subset of the possible interventions. In that case we would need to make modeling assumptions as to the nature of the interactions between the risk factors on survival (e.g. by assuming no interaction between risk factors on the mortality rate on a particular scale such as multiplicative or additive) both in order to obtain estimates of the $S_a(t)$ that are not too variable and to extrapolate to values of a outside the range of the data (i.e. to values of a to which no one was randomized). In the final analysis it is the area under the curves $S_a(t)$ that remains of interest. If our models are misspecified (e.g. we assume no interaction on an additive scale when in fact such interaction is present), the resulting estimates of $S_a(t)$ will be biased. Thus we would like to avoid use of models as much as possible. However the use of models cannot be done away with because of our inability to conduct a sufficiently large study.

In the randomized experiment of the previous paragraph, we can also estimate the effects of conditional (or dynamic) interventions. Let $L = (L^{(1)}, \dots, L^{(p)})$ denote a p -vector of measured baseline (pretreatment) covariates such as age, sex and measures of baseline health status. Let $d(l)$ be a function that assigns to each value of the vector L a value of a in the set \mathcal{A} of possible joint risk factor interventions. If a regime d assigns the same value a to each L , we refer to the regime d as non-dynamic. Otherwise, we refer to d as a conditional or dynamic treatment regime, strategy or plan as it individualizes the treatment (i.e. joint risk factor intervention) a subject receives based on the subject's value of L . A wise choice of d should allow us to optimize therapy for individuals and thus should be a better strategy than even the optimal non-dynamic intervention a^* discussed above. Let $S_d(t)$ be the survival curve that would result if we randomly assigned individuals to plan d . For subjects with a given value l of L , the conditional survival curve $S_d(t|L=l)$ given $L=l$ under regime d equals $S_a(t|L=l)$ for the value $a=d(l)$ that they receive under the plan. Thus for the population as a whole $S_d(t) = \sum_l S_a(t|L=l)pr(L=l)$ is weighted average of $S_a(t|L=l)$ with $a=d(l)$ and weights proportional to the fraction $pr(L=l)$ of the population with $L=l$. Thus the survival curve $S_d(t)$ of a dynamic regime can be estimated from the data in the randomized trial wherein each subject is randomized to a non-dynamic regime. Define $d_{op}(l)$ to be the treatment plan that minimizes the area under $S_d(t)$ over all possible dynamic and non-dynamic treatment plans d . The loss in life expectancy under intervention d compared to the optimal is the difference in area under $S_{d_{op}}(t)$ and $S_d(t)$. In an ideal world, we would estimate $d_{op}(l)$ from a large ideal

randomized study and, after analysing the trial, treat a new individual with $L = l$ with treatments $d_{op}(l)$.

3. SOME LIMITS OF IDEAL DATA

In this section, we return to the simple randomized trial of section 2.1.

3.1 EXTRAPOLATION

Although we have a precise estimate of the effect of this intervention on twenty year mortality of United States citizens in the calendar period 1983–2003, the trial provides no direct evidence concerning the following more relevant policy questions. What would be the continuing effect of this intervention on the United States population through 2013 or 2023? What would be the effect of this same intervention, if it began now in 2003 rather than in 1983? What would be the effect of a similar intervention on a population that differs from the United States population on both measured and unmeasured determinants of mortality including smoking, age, cholesterol, high blood pressure, lifestyle pattern, access to state-of-the-art health care, etc? Obviously any of these questions can only be addressed by assuming a model that extrapolates beyond the observed data. A common simple approach to extrapolation is to first statistically test from the available data whether the relative risk (equivalently, mortality ratio, hazard ratio, rate ratio) in the exposed compared to the non-exposed remains nearly constant both over the 20 years of follow-up and within levels of measured covariates such as age, ethnicity, socioeconomic status and smoking. If the test accepts the constant relative risk hypothesis then we extrapolate by assuming the same will be true if follow-up was continued past 2003, if the intervention was in 2003, and if the intervention was applied to a population with different smoking and risk factor distribution than the United States. In most studies, however, the power to detect deviations from a constant rate ratio is fairly small and there is rarely any strong biological reason to believe that rate ratios rather than other effect measures (such as rate differences) should be constant over time and location. Further, we have no way to test whether the effect of an intervention on a rate ratio scale is the same across groups that differ in unmeasured factors. Finally, even if we assume the relative risk to be constant, nevertheless, to estimate the effect of intervention on life expectancy, we still require a method to estimate covariate-calendar year-specific baseline rates in various populations in future years, since the effect on life expectancy depends both on the relative risk and on these baseline rates.

3.2 CONTAMINATION

In the discussion so far, we have assumed that the exposed do not “infect” or “contaminate” the unexposed with their exposure. This

assumption would not hold if subjects given, say, the aforementioned anti-smoking intervention distribute their anti-smoking materials to the control group. In that case, the beneficial effect of the intervention will be underestimated by the difference in the exposed and unexposed survival curves because a number of the unexposed will actually have been exposed. The difficulty is that the trial suffered from noncompliance in the sense that, contrary to our wishes, some of those assigned to no treatment actually received treatment. Certain analytic methods, referred to as instrumental variable methods, can partially adjust for this type of noncompliance if a randomly selected (possibly stratified) subsample of the trial population is interviewed to determine how much treatment they received as measured, say, by the fraction of the anti-smoking articles provided to the treatment group the controls actually read. This approach to correction for noncompliance is useful when the mechanism by which treatment exerts its effects is directly through access to the anti-smoking materials (Robins and Greenland 1996).

However, there are types of contamination that operate in quite different ways. For example, suppose that an effect of the treatment was that one of the treated subjects became so motivated that she started a public health campaign that resulted in the legal prohibition of smoking in public, leading to additional decreases in the cigarette consumption in both the exposed and unexposed. Then the difference in the treatment-arm specific survival curves would underestimate the total effect of the intervention (which should include the indirect effect through the legal prohibition) and correction by instrumental variable methods would not be possible. In this case a different design would be necessary. For example, one could use a cluster randomization design wherein different cities, counties or states are the experimental units and are randomly assigned as a whole to either treatment or control. The goal of the design is to have the experimental units sufficiently isolated from one another that one can be reasonably certain that between-unit contamination of treatment will not occur. If data from an appropriate cluster randomized design are not available, other different, and less reliable approaches to estimating the effect of the intervention on life expectancy must be used. Examples of such approaches include the following: (i) assume death rates would have remained unchanged had the intervention not occurred; (ii) specify and fit complex stochastic models for the mechanism by which the intervention reduced deaths, say, by assuming any decrease in the population exposure to cigarettes was wholly due to the intervention and modelling the effect on mortality of the observed decrease in smoking based on past or current data on changes in smoking and mortality; and (iii) create the observational equivalent of a cluster randomized design by assuming mortality and/or smoking data from other communities can be used to estimate what would have happened in the study population had no intervention taken place. A well known example of the type of contamination we are considering in this paragraph occurs in random-

ized trials studying the effect of vaccines on infectious disease and is the basis for the so-called phenomenon of “herd immunity” wherein an epidemic can be prevented in the unvaccinated (untreated) by vaccinating a sufficiently large fraction of the entire community.

The considerable difficulties caused by contamination and the need for extrapolation will not be considered further in this chapter due to space limitations and to the fact that the main goal of this chapter lies elsewhere. Manton et al. (1992) describe some models that may be used for extrapolation. We now turn to observational settings in which ideal epidemiologic data with which to estimate the effect of various treatments or risk factors even on the population under study may not be available.

4. OBSERVATIONAL DATA AND TIME-INDEPENDENT AND TIME-DEPENDENT INTERVENTIONS

For financial, ethical or logistical reasons, randomized trial evidence concerning the effectiveness of many if not most interventions is lacking and data from observational studies must be utilized. In this section we will use a hypothetical observational study of the effect of antiretroviral therapy on the progression of HIV-related disease as a specific example. It is well understood that causal effects can be estimated from observational studies only when data on all relevant time-independent and time-dependent confounding factors have been obtained. What is less well known is that for time-varying treatments, standard approaches to confounder control can be biased, even when the causal null hypothesis of no treatment effect is true and there are no unmeasured confounding factors. Specifically, the standard approach to the estimation of the causal effect of a time-varying treatment on survival has been to model the hazard of failure at t as a function of treatment history with a time-dependent proportional hazards model. Robins et al. (1986) have shown that, even in the absence of unmeasured confounding factors or model misspecification, the usual approach may be biased even under the causal null hypothesis, whether or not one further adjusts for the past history of measured covariates in the analysis, when (i) there exists a time-dependent risk factor (say CD4 lymphocyte count and/or *Pneumocystis carinii* pneumonia [PCP] history) for survival that also predicts subsequent treatment, and (ii) past treatment history predicts subsequent risk factor level. Specifically, condition (i) implies that the analysis that does not adjust for covariates is biased due to confounding by time-dependent risk factors such as CD4 count and/or PCP. Condition (ii) implies that the analysis that includes current CD4 count and/or PCP history as a regressor is biased since it adjusts for a variable (CD4 count and/or PCP history) affected by past treatment. In contrast to standard methods, estimation methods based on marginal structural models (MSMs), the parametric g-computation formula, and structural nested models provide consistent estimates of causal effects whenever unmeasured confounding

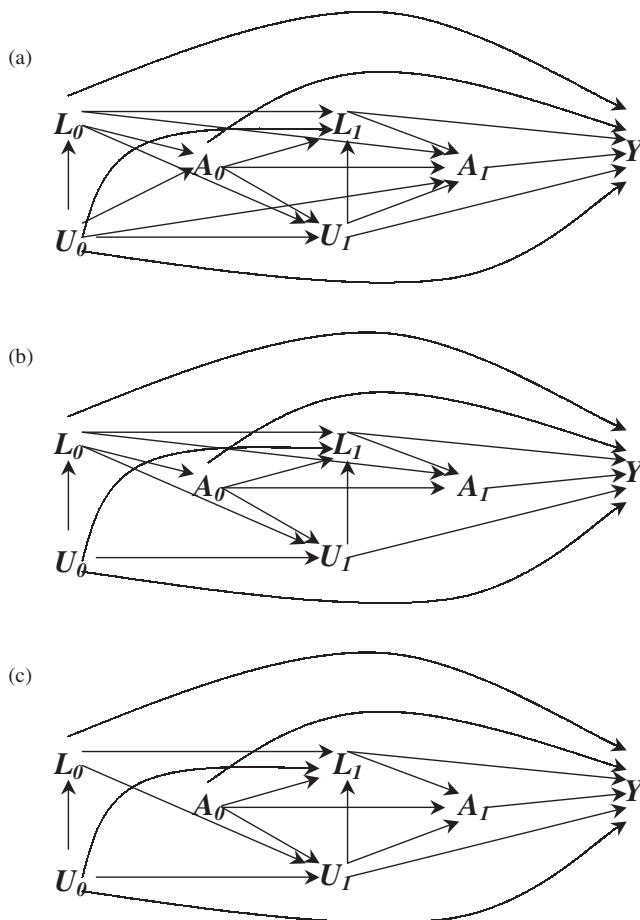
and model misspecification are absent. (See the article, discussion, and rejoinder in Robins et al. (1999) for additional details.) To describe difficulties with standard approaches and the basis for the effectiveness of these three alternatives, it will be useful to informally introduce causal directed acyclic graphs (DAGs) as discussed by Spirtes et al. (1993), Pearl (1995), and Greenland et al. (1999).

A causal graph is a DAG in which the vertices (nodes) of the graph represent variables, the directed edges (arrows) represent direct causal relations between variables, and there are no directed cycles, since no variable can cause itself. For a DAG to be causal, the variables represented on the graph must include the variables and additional unmeasured variables, such that if any two measured variables on the graph have a cause in common, that common cause is itself included as a variable on the graph.

A variable is a cause of a second variable if there is a directed path from the first variable to the second consisting solely of arrows pointing towards the second variable.

As an example consider a follow-up study of AIDS patients. Let $A(t) = A_t$ be the dose of the treatment or exposure of interest, say an antiretroviral therapy, at t with time measured in months since start of follow-up. The time t at which a treatment occurs will either be placed in parentheses or subscripted, depending on the context. Let Y be a dichotomous outcome of interest (e.g. $Y = 1$ if HIV RNA is not detectable in the blood and is 0 otherwise) measured at end of follow-up at time $K + 1$. Our goal is to estimate the causal effect of the time-dependent treatment $A(t)$ on the outcome Y .

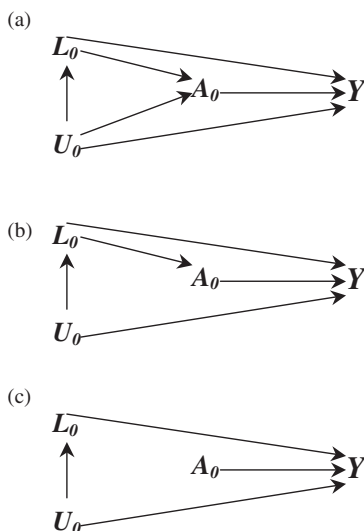
Figure 28.1 is a causal graph that represents our study with $K = 1$. In Figure 28.1, $L(t) = L_t$ represents the value at time t of the vector of all measured causal risk factors for the outcome, such as CD4 count, white blood count (WBC), red blood count (RBC), the presence or absence of various opportunistic infections such as PCP, age and weight. We assume that $L(t)$ is temporally prior to $A(t)$ since physicians commonly obtained data recorded in $L(t)$ such as CD4 count before deciding on a treatment $A(t)$ to be given in month t . Similarly, $U(t) = U_t$ represents the value at time t of all unmeasured causal risk factors for Y . Figure 28.1(b) differs from Figure 28.1(a) only in that the arrows from the unmeasured risk factors into the treatment variables have been removed. When, as in Figure 28.1(b), there are no arrows from unmeasured risk factors directly into treatment variables, we say that there are no unmeasured confounders given data on the measured confounders $L(t)$. Figure 28.1(c) differs from Figures 28.1(a) and 28.1(b) in that none of the causal risk factors for Y (measured or unmeasured) has arrows into any treatment variable. Note, however, that earlier treatment $A(0)$ can causally affect later treatment $A(1)$. When, as in Figure 28.1(c), there are no arrows from any (non-treatment) risk factor into any treatment variable, there is no confounding by either measured or unmeasured factors, in which

Figure 28.1 Causal graphs for a time-dependent exposure

case we say that treatment is unconfounded or, equivalently, causally exogenous.

The distinctions drawn above apply equally to more familiar point treatment studies where the treatment is not time-dependent. As indicated in Figure 28.2, a point treatment study is a special case of the general set-up in which $K = 0$. Figures 28.2(a)–28.2(c) are the analogues of Figures 28.1(a)–28.1(c) for a point treatment study.

Our causal DAGs would be of no use without an assumption linking the causal structure represented by the DAG to the statistical data obtained in an epidemiologic study. Recall that if a set of variables X is statistically independent of (i.e. unassociated with) another set of variables Y conditional on a third set of variables Z , then within joint

Figure 28.2 Causal graphs for a point exposure A_0 

strata defined by the variables in Z , any variable in X is unassociated with any variable in Y . For example, suppose all variables are dichotomous and the set Z consists of the two variables Z_1 and Z_2 . Then conditional independence implies that the population odds ratio and population risk ratio between any variable in X and any variable in Y is 1 within each of the $4 = 2^2$ strata of Z : $(Z_1, Z_2) = (0, 0)$, $(Z_1, Z_2) = (0, 1)$, $(Z_1, Z_2) = (1, 0)$ and $(Z_1, Z_2) = (1, 1)$. We use the symbol $\perp\!\!\!\perp$ to indicate statistical independence, e.g. $X \perp\!\!\!\perp Y | Z$ means X is conditionally independent of Y given Z . The following so-called causal Markov assumption links the causal structure of the DAG with the various statistical independencies.

4.1 CAUSAL MARKOV ASSUMPTION (CMA)

On a causal graph any variable that is not caused by a given variable V will be independent of (i.e. unassociated with) V conditional on (i.e. within joint strata defined by) V 's direct causes.

Thus, in Figure 28.1(c), $A(t)$ being causally exogenous implies that $A(0) \perp\!\!\!\perp \{L(0), U(0)\}$ and $A(1) \perp\!\!\!\perp \{L(0), L(1), U(0), U(1)\} | A(0)$.

As in any observational study, we cannot determine from the observed data whether there is confounding by unmeasured risk factors. We can only hope that whatever residual confounding there may be due to the $U_1 < y$ is small. However, as discussed further below, under the untestable assumption that there is no unmeasured confounding given the L_k , we

can empirically test from the data whether treatment is causally exogenous. Specifically, a sufficient condition for treatment to be unconfounded is that, at each time k , among subjects with the same past treatment history A_0, \dots, A_{k-1} , the treatment A_k is unassociated with (i.e. statistically independent of) the past history of measured covariates L_0, \dots, L_k . In particular, in our point treatment study, treatment will be causally exogenous if A_0 is unassociated with L_0 .

4.2 INVERSE PROBABILITY OF TREATMENT WEIGHTED ESTIMATION

In this section, our goal is to estimate using MSMs the causal effect of the time-dependent treatment $A(t)$ on the mean of Y , which for a dichotomous $(0, 1)$ response is just the probability that $Y = 1$. In this section, we assume that there is no loss to follow-up so Y is observed on each study subject. For any time-dependent variable we use overbars to denote the history of that variable up to and including t . For example, $\bar{L}(t) = [L(0), L(1), L(2), \dots, L(t)]$ is the covariate process through t . Consider first the association (i.e. regression) model that states that the mean of Y , given treatment history $\bar{A} \equiv \bar{A}(K)$, is a linear logistic function of a subject's duration of antiretroviral therapy. That is,

$$E[Y | \bar{A}] = g(\bar{A}; \gamma)$$

where

$$g(\bar{A}; \gamma) = \frac{\exp[\gamma_0 + \gamma_1 Dur(\bar{A})]}{1 + \exp[\gamma_0 + \gamma_1 Dur(\bar{A})]} \quad (1)$$

$Dur(\bar{A}) = \sum_{k=0}^K A(k)$ is the subject's duration of treatment in months, and $A(k)$ equals 1 if the subject is on treatment in month k and is 0 otherwise. That is, we are assuming a linear logistic regression model

$$\text{logit } pr(Y = 1 | \bar{A}) = \gamma_0 + \gamma_1 Dur(\bar{A}) \quad (2)$$

The logistic regression maximum likelihood estimator (MLE) of $\gamma = (\gamma_0, \gamma_1)$ that is computed by all standard packages maximizes $\prod_{i=1}^n Lik_i(\gamma)$ with $Lik_i(\gamma) = g(\bar{A}_i; \gamma)^{y_i} [1 - g(\bar{A}_i; \gamma)]^{1-y_i}$ being the likelihood contribution for a single subject and n the sample size.

Assuming the association model (1) is correct, when does γ_1 and γ_2 have a causal interpretation? To answer this question, imagine that the decision to administer treatment at each time t was made totally at random by the treating physician. In that hypothetical randomized trial, treatment at time t is not expected to be associated with the history up

to t of any measured or unmeasured prognostic factors (i.e. there is no confounding). In the absence of confounding, association implies causation and we would expect γ_1 to represent the effect of antiretroviral therapy on the mean of Y . More generally, the marginal association between treatment and response represents causation whenever the treatment is causally exogenous, that is, the conditional probability of receiving a treatment $A(t)$ at time t given past treatment and prognostic factor history for Y (measured and unmeasured) depends only on past history of treatment $\bar{A}(t-1)$ as in Figures 28.1(c) and 28.2(c). A more technical definition is provided below after we define counterfactual outcomes. It is well-recognized in the social sciences, econometrics, epidemiologic, and biostatistical literature that the treatment parameters of a correctly specified association model will have a causal interpretation if treatment is causally exogenous.

To help assess whether antiretroviral therapy may be causally exogenous, we introduce the concept of “statistical exogeneity”. We say that treatment $A(t)$ is a “statistically exogenous or ancillary” process if the probability of receiving treatment at time t does not depend on the history of measured time-dependent prognostic factors up to t , conditional on treatment history prior to t , i.e.

$$\bar{L}(t) \perp\!\!\!\perp A(t) \mid \bar{A}(t-1)$$

Note that a nearly necessary condition for $A(t)$ to be “causally exogenous” is for it to be “statistically exogenous”. However, that a process is “statistically exogenous” does not imply it is “causally exogenous”, because there may be unmeasured prognostic factors $\bar{U}(t)$ for the outcome (i.e. confounders) that predict the probability of treatment $A(t)$ at time t given past treatment history. Thus we can test from the data whether $A(t)$ is statistically exogenous as it is a relationship between observed variables, but are unable to test whether a statistically exogenous process is causally exogenous. However, as mentioned above and discussed further below, if we make the untestable assumption that there are no unmeasured confounders, then statistical exogeneity will imply causal exogeneity.

Suppose $A(t)$ is discrete and we can correctly model both the probability $f[a(t) \mid \bar{l}(t), \bar{a}(t-1)]$ of taking treatment $a(t)$ on day t as a function of past treatment $\bar{a}(t-1)$ and measured prognostic factor history $\bar{l}(t)$, and the probability $f[a(t) \mid \bar{a}(t-1)]$ of taking treatment $a(t)$ in month t as a function only of past treatment $\bar{a}(t-1)$ history. Here we use the convention that random variables (i.e. variables whose values can differ from subject to subject) are denoted by upper case letters. Lower case letters denote possible values of the corresponding random variables. Thus, for example, $f[a(t) \mid \bar{a}(t-1)]$ is the proportion of subjects in the target population with treatment $A(t)$ equal to $a(t)$ among subjects with past treatment history $\bar{A}(t-1)$ equal to $\bar{a}(t-1)$. We could measure the degree to

which the treatment process is statistically non-exogenous through time t by the time t -specific random variable

$$SW(t) = \prod_{k=0}^t \frac{f[A(k)|\bar{A}(k-1)]}{f[A(k)|\bar{A}(k-1), \bar{L}(k)]} \quad (3)$$

where $\prod_{k=0}^t z(k)$ is the product $z(0) \times z(1) \times \dots \times z(t)$, $f[A(k)|\bar{A}(k-1), \bar{L}(k)]$ is, by definition, the conditional probability mass function $f[a(k)|\bar{a}(k-1), \bar{l}(k)]$ with $(a(k), \bar{a}(k-1), \bar{l}(k))$ evaluated at a subject's data $(A(k), \bar{A}(k-1), \bar{L}(k))$ and $f[A(k)|\bar{A}(k-1)]$.

For example, if for a given subject $A(k)$ is zero and there are 55 other subjects with the same $\bar{A}(k-1), \bar{L}(k)$ history of whom 25 have $A(k)$ of zero and 70 subjects with the same $\bar{A}(k-1)$ history of whom 32 have $A(k)$ of zero, then $f[A(k)|\bar{A}(k-1), \bar{L}(k)]$ is $20/55$ and $f[A(k)|\bar{A}(k-1)]$ is $32/70$ for the subject. Informally, the denominator in each term in $SW(t)$ is the probability that a subject received his own observed treatment, $A(k)$, at time k given his past antiretroviral treatment and measured prognostic factor history. Informally, the numerator is the probability that a subject received his observed treatment conditional on his past antiretroviral treatment history, but not further adjusting for his past prognostic factor history. Note that the numerator and denominator of $SW(t)$ are equal for all t with probability 1 if and only if the treatment process is statistically exogenous, i.e. $\bar{L}(t) \perp\!\!\!\perp A(t) | \bar{A}(t-1)$. In practice $SW(t)$ will have to be estimated from the data but, for pedagogical purposes, assume for now that it is known.

When $A(t)$ is statistically non-exogenous, we shall consider estimating γ by a weighted logistic regression in which a subject is given the weight $SW \equiv SW(K)$. Standard software packages for logistic regression will allow the user to specify the subject-specific weight SW . The weighted logistic regression estimator, which we will refer to as an inverse-probability-of-treatment-weighted (IPTW) estimator, is the

maximizer of $\prod_{i=0}^n [Lik_i(\gamma)]^{SW_i}$. This weighted logistic regression would

agree with the usual unweighted analysis described above just in the case in which $A(t)$ were statistically exogenous. The IPTW estimator is an extension to longitudinal causal inference models of estimators proposed by Horvitz and Thompson (1952).

If the vector of prognostic factors recorded in $\bar{L}(t)$ constitutes all relevant time-dependent prognostic factors (i.e. confounders) so that there are no unmeasured confounders (as in Figures 28.1(b) or 28.2(b)), then, whether or not the treatment process is statistically exogenous, the

weighted logistic regression estimator of γ_1 will converge to a quantity β_1 that can be interpreted as the causal effect of antiretroviral therapy on the mean of Y (on the log odds ratio scale). In contrast, when $A(t)$ is statistically non exogenous, the usual unweighted logistic regression estimator will still converge to γ_1 , but now γ_1 will have no causal interpretation. We now give a formal mathematical meaning to the informal concepts of the causal effect of antiretroviral therapy on the mean of Y .

4.3 COUNTERFACTUALS AND MARGINAL STRUCTURAL MODELS

To formalize our results, we use counterfactual or potential outcomes. Neyman (1923) introduced counterfactual outcomes to analyse the causal effect of time-independent treatments in randomized experiments. Rubin (1978) championed Neyman's idea and emphasized the usefulness of counterfactuals in the analysis of the causal effects of time-independent treatments from observational data. Robins (1986, 1987) proposed a formal counterfactual theory of causal inference that extended Neyman's (1923) time-independent treatment theory to longitudinal studies with both direct and indirect effects and sequential time-varying treatments and confounders. In this theory, for any fixed history of antiretroviral therapy \bar{a} , $Y_{\bar{a}}$ is defined to be the random variable representing a subject's outcome had, possibly contrary to fact, the subject been treated with \bar{a} rather than his observed treatment \bar{A} . Note that \bar{a} is a possible value of the random variable \bar{A} . For each possible history \bar{a} we are assuming a subject's response $Y_{\bar{a}}$ is well defined, although generally unobserved. Indeed we only observe $Y_{\bar{a}}$ for that treatment history \bar{a} equal to a subject's actual treatment history \bar{A} , i.e. a subject's observed outcome Y equals $Y_{\bar{A}}$. This identity is the fundamental "consistency" assumption that links the counterfactual data $Y_{\bar{a}}$ to the observed data (Y, \bar{A}) .

Note that if, at each time t , $A(t)$ can take but one of two values (0 for untreated and 1 for treated) and the study duration is K months, then there are 2^K different $Y_{\bar{a}}$ values associated with each subject as there are 2^K possible treatment patterns only one of which is actually observed for a given subject. Then, formally, the statement that the effect of treatment history on the mean of Y is a linear logistic function of duration of antiretroviral therapy is the statement that, for each \bar{a} ,

$$E[Y_{\bar{a}}] = g[\bar{a}; \beta]$$

where

$$g[\bar{a}; \beta] = \frac{\exp(\beta_0 + \beta_1 Dur(\bar{a}))}{1 + \exp(\beta_0 + \beta_1 Dur(\bar{a}))} \quad (4)$$

$\beta = (\beta_0, \beta_1)$, and $Dur(\bar{a}) = \sum_{k=0}^K a(k)$ is the subject's duration of treatment

under the treatment history \bar{a} . We refer to this model as a MSM for the effect of antiretroviral therapy on the mean of Y , since it is a model for the marginal distribution of counterfactual variables and, in the econometric and social science literature, causal models (i.e. models for counterfactual variables) are often referred to as structural.

The parameter β of our MSM encodes the magnitude of the average causal effects of the treatment on the outcome. By definition, the causal effect of treatment regime \bar{a} on the outcome Y for a given study subject is the difference $Y_{\bar{a}} - Y_{\bar{0}}$ between his outcome $Y_{\bar{a}}$ when treated with regime \bar{a} and his outcome $Y_{\bar{0}}$ when never treated. Thus the average causal effect of regime \bar{a} is $E[Y_{\bar{a}} - Y_{\bar{0}}] = E[Y_{\bar{a}}] - E[Y_{\bar{0}}] = g(\bar{a}; \beta) - g(\bar{0}; \beta)$, which depends on β . If β_1 is zero, we say that there is no effect of treatment \bar{a} on the outcome since $E[Y_{\bar{a}}] - E[Y_{\bar{0}}]$ is the same for all \bar{a} . In contrast, the association parameter γ_1 lacks a causal interpretation when treatment is not causally exogenous. Furthermore, the optimal non-dynamic intervention \bar{a}^* is the value of \bar{a} for which $E[Y_{\bar{a}}] = g(\bar{a}; \beta)$ is the greatest if our goal is to maximize the probability that HIV will not be detected in the serum.

4.4 FORMAL DEFINITIONS OF CAUSAL EXOGENEITY AND NO UNMEASURED CONFOUNDERS

We are now in a position to offer mathematically precise definitions of causal exogeneity and of no unmeasured confounders. *Let $\{Y_{\bar{a}}\}$ be the set of all counterfactual outcomes $Y_{\bar{a}}$ as \bar{a} varies.* Formally, we say that the treatment process $A(t)$ is causally exogenous if,

$$\{Y_{\bar{a}}\} \perp\!\!\!\perp A(t) \mid \bar{A}(t-1) \quad (5)$$

which is mathematically equivalent to the statement that $\{Y_{\bar{a}}\}$ is independent of \bar{A} . Note that even when $A(t)$ is "causally exogenous", if the treatment has an effect on the outcome, then the observed outcome $Y = Y_{\bar{A}}$ will not be independent of \bar{A} , since $Y_{\bar{A}}$ is a function of a subject's observed treatment history \bar{A} itself. Given the covariates recorded in $L(t)$, following Robins (1987) we say there are no unmeasured confounders for the effect of $A(t)$ on Y if

$$\{Y_{\bar{a}}\} \perp\!\!\!\perp A(t) \mid \bar{A}(t-1), \bar{L}(t) \quad (6)$$

We shall also refer to the assumption of no unmeasured confounders as the assumption that treatment $A(t)$ is sequentially randomized given the past. This assumption generalizes Rosenbaum and Rubin's (1983) assumption of ignorable treatment assignment to longitudinal studies with time-varying treatments and confounders. The assumption states

that, conditional on treatment history and the history of all recorded covariates up to t , treatment at t is independent of the counterfactual random variables $Y_{\bar{a}}$. This will be true if all prognostic factors for (i.e. predictors of) Y that are used by physicians to determine whether treatment is given at t are recorded in $\bar{L}(t)$ and $\bar{A}(t-1)$. That is, as in Figure 28.1(b), the causal graph generating the data has no arrows directly from any unmeasured causal risk factors for Y directly into treatment. For example, since physicians tend to administer prophylaxis to subjects with previous bouts of PCP, and in untreated subjects PCP predicts Y , the assumption of no unmeasured confounders would be false if $\bar{L}(t)$ does not contain PCP history. It is the primary goal of the epidemiologists conducting an observational study to collect data on a sufficient number of covariates to ensure that the assumption of no unmeasured confounders will be at least approximately true.

In an observational study, the assumption of no unmeasured confounders cannot be guaranteed to hold even approximately and it is not subject to empirical test. Therefore, it may be useful to investigate the sensitivity to violations of the assumption through a formal sensitivity analysis. Robins et al. (1999, rejoinder) and Robins et al. (1999) provide details.

Robins (1999) proved that when there are no unmeasured confounders, (i) statistical exogeneity implies causal exogeneity, (ii) the weighted logistic regression estimator using the weights SW converges to the parameter β of the MSM (4) for $E[Y_{\bar{a}}]$, and (iii) the probability limit γ of the usual unweighted logistic estimator generally differs from the causal parameter β of the MSM unless the treatment process is statistically exogenous. Here we provide an informal heuristic argument for (ii). View each person as a member of a pseudo-population consisting of SW copies of themselves. In this new pseudo-population, it can be shown that $\bar{L}(t)$ does not predict treatment at t given past treatment history, and thus we have created a pseudo-population in which treatment is causally exogenous. Furthermore, the causal effect of treatment on Y in the pseudo-population is the same as in the original population. That is, if $E[Y_{\bar{a}}] = g(\bar{a}; \beta)$ in the true population, the same will be true of the pseudo-population. Hence, we would like to do ordinary logistic regression in the pseudo-population. But that is what our weighted logistic regression estimator is doing, since the weights create, as required, SW copies of each subject.

We can generalize our MSM (4) slightly and model the marginal distribution of $Y_{\bar{a}}$ within levels of a subset V of the pre-treatment (baseline) covariates $L(0)$. Then, our marginal structural logistic model (4) could be modified to

$$E[Y_{\bar{a}}|V] = \frac{\exp(\beta_0 + \beta_1 Dur(\bar{a}) + \beta_2'V + \beta_3' Dur(\bar{a})V)}{1 + \exp(\beta_0 + \beta_1 Dur(\bar{a}) + \beta_2'V + \beta_3' Dur(\bar{a})V)}$$

β'_3 measures how the magnitude of the effect of $Dur(\bar{a})$ is modified by the pretreatment covariates V . An IPTW estimator of the parameter β can be obtained by weighted logistic regression with weights SW except now the logistic model includes $Dur(\bar{A})$ and V as regressors, and $SW(t)$ is redefined to be

$$SW(t) = \prod_{k=0}^t \frac{f[A(k)|\bar{A}(k-1), V]}{f[A(k)|\bar{A}(k-1), \bar{L}(k)]} \quad (7)$$

Note V is already included in the denominator, since V is a subset of the variables in $L(0)$.

Let $d(V)$ be a function (regime) that assigns to each value of the vector V a value of \bar{a} in the set V of possible interventions. If a regime d assigns the same value \bar{a} to each v we refer to the regime d as non-dynamic. Otherwise, we refer to d as a conditional or baseline-dynamic treatment regime, strategy or plan as it individualizes the treatment history a subject receives based on the subject's value of the baseline variables recorded in V , a subset of $L(0)$. A wise choice of d should allow us to optimize therapy for individuals and thus should be a better strategy than even the optimal non-dynamic intervention a^* . Let $E[Y_d]$ be the probability of being without HIV in the serum if all subjects followed plan d . For subjects with a given value v of V , the conditional expectation $E[Y_d|V=v]$ given $V=v$ under regime d equals $E[Y_{\bar{a}}|V=v]$ for the value $\bar{a} = d(v)$ that they receive under the plan. Thus for the population as a whole $E[Y_d] = \sum_v E[Y_{\bar{a}}|V=v]pr(V=v)$ is a weighted average of $E[Y_{\bar{a}}|V=v]$ with $a = d(v)$ and weights proportional to the fraction $pr(V=v)$ of the population with $V=v$. Then $d_{op}^{baseline}(v)$ is the treatment plan that maximizes $E[Y_d]$ over all possible baseline-dynamic and non-dynamic treatment plans d . Now even $d_{op}^{baseline}(v)$ only allows one to optimize treatment history \bar{a} based on pretreatment (baseline) variables. However, with time-varying treatments it is usually important to dynamically choose the treatment at each time t based on a subject's entire covariate history up to time t . For example, consider drug treatment for a chronic disease. When a drug becomes toxic to a subject, the optimal strategy is to stop the drug (or reduce the dose) at least temporarily. One cannot know when to stop the drug based on baseline covariates. Rather, the optimal treatment strategy must allow treatment decisions to be based on a subject's evolving covariate history. The best methods for estimating the effects of true dynamic regimes are not based on MSMs. Further discussion is provided in section 5.

4.5 MARGINAL STRUCTURAL COX PROPORTIONAL HAZARDS MODEL

MSMs can easily be extended to failure time outcomes by specifying a marginal structural Cox proportional hazards model such as

$$\lambda_{T\bar{a}}(t|V) = \lambda_0(t) \exp(\beta_1 a(t) + \beta_2' V + \beta_3' a(t)V), \quad (8)$$

where $T_{\bar{a}}$ is the subject's time to death if he had followed anti-retroviral therapy history \bar{a} , $\lambda_{T_{\bar{a}}}(t|V)$ is the hazard (force of mortality) of $T_{\bar{a}}$ at t conditional on having pretreatment variables V , $\lambda_0(t)$ is an unspecified baseline hazard function, $\exp(\beta_1 + \beta'_3 V)$ is the causal rate ratio for the effects of treatment at level V of a vector of baseline regressors including age, calendar year, CD4 count, CD8 count, WBC count, RBC count, platelets, etc. For variety, we have chosen a model which specifies that the hazard of death at time t depends on current treatment status rather than the duration of treatment. Other dose-response models could be used.

Let T be a subject's observed failure (i.e. death) time, so that $T = T_{\bar{a}}$. Arguing as above, Robins (1999) shows that, in the absence of censoring, a consistent estimator of the unknown parameter $\beta = (\beta_1, \beta'_2, \beta'_3)'$ is obtained by fitting the ordinary time-dependent Cox model

$$\lambda_T(t|\bar{A}(t), V) = \lambda_0(t) \exp(\gamma_1 A(t) + \gamma'_2 V + \gamma'_3 A(t)V) \quad (9)$$

except that the contribution of a subject to a calculation performed on a subject i at risk at time t is weighted by $SW_i(t)$, as defined in (7) with $T > k$ added to the conditioning event. Note the subject-specific weights change with time. Few standard Cox proportional hazards software programs allow for time-varying weights. To avoid this software problem one can fit a weighted pooled logistic regression treating each person-month as an observation and allowing for a time-dependent intercept. That is, one can fit, by weighted logistic regression using weights $SW(t)$, the model

$$\begin{aligned} \text{logit } pr[D(t) = 1 | D(t-1) = 0, \bar{A}(t-1), V] \\ = \gamma_0(t) + \gamma_1 A(t-1) + \gamma'_2 V + \gamma'_3 A(t-1)V \end{aligned} \quad (10)$$

where $D(t) = 0$ if a subject was alive at time t and 1 if the subject died at month t , and $\gamma_0(t)$ is a time-specific intercept. This method offers the advantage of being easily programmed in any standard statistical package. Under our assumptions we thereby obtain a consistent estimator of the parameter vector β of the MSM

$$\begin{aligned} \text{logit } pr[D_{\bar{a}}(t) = 1 | D_{\bar{a}}(t-1) = 0, V] \\ = \beta_0(t) + \beta_1 a(t-1) + \beta'_2 V + \beta'_3 a(t-1)V \end{aligned} \quad (11)$$

When the death rate in any given month t is small, the parameters of (11) and (8) closely approximate one another.

Because of the weights, the standard error estimates outputted by a standard logistic program are invalid and may be either too large or too small. To overcome this difficulty, model (10) should be fit using a generalized estimating equations (GEE) (Liang and Zeger 1986) program

which outputs robust variance estimators. The robust variance GEE estimators provide a conservative confidence interval for the β (Robins, 1999). That is, the 95% Wald confidence interval calculated as $\beta \pm 1.96 \times (\text{robust})$ standard error is guaranteed to cover the true β at least 95% of the time in large samples.

We now describe how to accommodate censoring in the analysis. We defined a subject as right censored at time t (i.e. $C(t) = 1$) if by time t he either dropped out of the study or reached administrative end of follow-up alive.

We say that censoring is ignorable or non-informative if the conditional cause-specific hazard of being censored at k among subjects alive and uncensored up to k does not depend on the failure times $T_{\bar{a}}$ given $\bar{A}(k-1)$, and the time-dependent covariate $\bar{L}(k-1)$ history prior to k . Under the assumptions of ignorable censoring and no unmeasured confounding, Robins (1999) shows that we still obtain from fitting (10) consistent estimators of β if we weight a subject alive and uncensored at month t by $SW(t) \times SW^\dagger(t)$ where (i)

$$SW^\dagger(t) = \prod_{k=0}^t \frac{\text{pr}[C(k) = 0 | \bar{C}(k-1) = 0, \bar{A}(k-1), V, T > k]}{\text{pr}[C(k) = 0 | \bar{C}(k-1) = 0, \bar{A}(k-1), \bar{L}(k), T > k]}$$

is informally the inverse of the ratio of a subject's probability of remaining uncensored up to month t divided by that probability calculated as if there had been no time-dependent determinants of censoring except past treatment history and V , and (ii) we modify our definition (7) of $SW(t)$ to add $C(k) = 0$ to the conditioning events both in the numerator and the denominator. The denominator of $SW(t) \times SW^\dagger(t)$ is informally the probability that a subject would have his own observed treatment and censoring history through month t .

Hernan et al. (2000) describe how to estimate the weights $SW(t)(SW^\dagger(t))$ from the data using pooled logistic regression models with treatment (censoring) at each time k as the response variable. Substituting out estimated weights into our IPTW estimators allows us to estimate, under the assumption of no unmeasured confounders, logit $\text{pr}[D_{\bar{a}}(t) = 1 | D_{\bar{a}}(t-1) = 0, V] = \beta_0(t) + \beta_1 a(t-1) + \beta_2 V + \beta_3 a(t-1)V$ and

thus the conditional survival curves $S_{\bar{a}}(t|V) = \prod_{k=0}^t \text{pr}[D_{\bar{a}}(k) = 0 | D_{\bar{a}}(k-1) = 0, V]$ that would be observed if all subjects have followed regime \bar{a} .

Again let $d(V)$ be a function (regime) that assigns to each value of the baseline vector V a value of \bar{a} in the set V of possible interventions. Let $S_d(t)$ be the survival curve if all subjects followed plan d . For subjects with a given value v of V , the conditional survival curve $S_d(t|V = v)$ given $V = v$ under regime d is $\sum_v S_{\bar{a}}(t|V = v) \text{pr}(V = v)$ is a weighted average of $S_{\bar{a}}(t|V = v)$ with $a = d(v)$ and weights proportional to the fraction $\text{pr}(V = v)$ of the population with $V = v$. Then $d_{op}^{baseline}(v)$ is the treatment

plan that minimizes the area under $S_d(t)$ over all possible baseline-dynamic and non-dynamic treatment plans d . Now even $d_{op}^{baseline}(v)$ only allows one to optimize treatment history \bar{a} based on pretreatment (baseline) variables. The best methods for estimating the effects of general dynamic regimes are not based on MSMs. Further discussion is provided in section 5.

We have seen that under the assumption of no unmeasured confounders, IPTW estimation of a marginal structural Cox proportional hazards model can, in contrast with standard methods, be used to estimate the effect of time-varying treatments on survival.

The correctness of the resulting causal inferences is dependent on three key assumptions. First, we must assume that the covariates in $L(t)$ are sufficient to adjust for both confounding and for selection bias due to loss to follow-up. This implies that we have available, in each month, data recorded in $L(t)$ on the history of all time-dependent covariates that (i) are independent predictors of death and (ii) independently predict the probability of changes in treatment and/or of being censored in that month. As in all observational studies, this fundamental assumption cannot be empirically tested. In practice, this would never be precisely or sometimes even approximately true. As described earlier, methods have recently been developed which allow one to evaluate the sensitivity of one's estimates to increasing violation of this assumption.

Second, we must assume that our models for changes in treatment and censoring, given past covariate and treatment history, are correctly specified. Last, we need to assume that our MSM for the effect of antiretroviral therapy on mortality is correctly specified.

Even when estimating the effect of a time-independent treatment using standard statistical models, the same assumptions (no unmeasured confounders, non informative censoring, and no model misspecification) are required to endow the parameters with a causal interpretation. Furthermore, when estimating the effect of a time-varying treatment, our assumptions are less restrictive than those required by standard analyses: an approach based on IPTW estimation of MSMs does not require for validity the absence of confounding by time-dependent covariates.

5. ALTERNATIVES TO MSMs

Before introducing MSMs, Robins and coauthors introduced two other methods for estimation of the causal effect of a time-varying treatment in the presence of time-varying confounders: the parametric g-computation algorithm formula estimator (Robins 1986), and g-estimation of structural nested models (Robins et al. 1992). When (i) both treatment and the confounders are discrete variables, (ii) they are measured at only a few time points, and (iii) the sample size is large, then

estimation can be carried out using fully saturated models (i.e. non-parametrically) and all three methods are precisely equivalent. They differ when, as in observational studies with sparse multivariate data, one must introduce modelling assumptions.

A major advantage of MSMs is that they resemble standard models. For example, the logistic MSM and the Cox proportional hazards MSM described above are the natural way to extend the ordinary logistic and time-dependent Cox models to allow for estimation of causal effects of time-dependent treatments.

However a major advantage of the parametric g-computation algorithm formula estimator and g-estimation of structural nested models over MSMs is that these models are much more useful than MSMs for estimating both interactions between treatment and time-dependent covariates and the effect of dynamic treatment regimes (Robins 1999). Due to space limitations, we will focus in this chapter on the parametric g-computation algorithm formula estimator, and we will not consider g-estimation of structural nested models.

Now for any variable Z , let \mathcal{Z} be the support (i.e. the possible values) of Z . Define a treatment regime or plan d to be a collection of $K + 1$ functions $d = \{d_0, \dots, d_K\}$ where $d_m: \bar{\mathcal{L}}_m \rightarrow \mathcal{A}_m$ maps histories $\bar{l}_m \in \bar{\mathcal{L}}_m$ into a treatment $d_m(\bar{l}_m) \in \mathcal{A}_m$. If $d_m(\bar{l}_m)$ is a constant, say a_m , not depending on \bar{l}_m for each m , we say regime d is non-dynamic and write $d = \bar{a}$, $\bar{a} \equiv (a_0, \dots, a_K)$. Otherwise, d is dynamic. We let \mathcal{D} be the set of all regimes d . Let $f(o)$ and $F(o)$ represent the density and distribution function of the observed data $O = (\bar{A}_K, \bar{L}_{K+1})$.

Associated with each regime d is the distribution $F_d(o)$ with density $f_d(o)$ that represents the distribution of the observed data had, contrary to fact, all subjects in the population been treated with regime d . Suppose the assumption of no unmeasured confounders holds for the joint outcomes $\bar{L}_{\bar{a}} = \bar{L}_{\bar{a}, K+1}$ [i.e. Equation 6 holds with $\{\bar{L}_{\bar{a}}\}$ replacing $\{\bar{Y}_{\bar{a}}\}$]. Then given a regime $d = (d_0, d_1, \dots, d_K)$ and the joint density

$$f(o) = f(l_0) \times f(a_0|l_0) \times \dots \times f(a_K|\bar{l}_K, \bar{a}_{K-1}) \times f(l_{K+1}|\bar{l}_K, \bar{a}_K), \quad (12)$$

of the observed data (written as a Markov factorization in temporal order), $f_d(o)$ is the density $f(o)$ except that, in the factorization (12), $f(a_0|l_0)$ is replaced by a degenerate distribution at $a_0 = d_0(l_0)$, $f(a_1|l_1, a_0, l_0)$ is replaced by a degenerate distribution at $a_1 = d_1(l_0, l_1)$, and, in general, $f(a_k|\bar{l}_k, \bar{a}_{k-1})$ is replaced by a degenerate distribution at $a_k = d_k(\bar{l}_k)$. That is the density $f_d(o)$ is given by

$$f_d(o) = f(l_0) \times f_d(a_0|l_0) \times \dots \times f_d(a_K|\bar{l}_K, \bar{a}_{K-1}) \times f(l_{K+1}|\bar{l}_K, \bar{a}_K), \quad (13)$$

where $f_d(a_k|\bar{l}_k, \bar{a}_{k-1})$ is equal to 1 if $a_k = d_k(\bar{l}_k)$ and $f_d(a_k|\bar{l}_k, \bar{a}_{k-1})$ is equal to 0 otherwise.

Again suppose the outcome of interest is L_{K+1} which is assumed to be univariate and shall be denoted by Y . In the following, let $d(\bar{l}_k) \equiv (d_0(\bar{l}_0), \dots, d_k(\bar{l}_k))$ and $d_k(\bar{l}_k)$ denote values of \bar{A}_k and A_k respectively. Then the chance of having history (y, \bar{l}_K) under the distribution $F_d(o)$ is $f_d(y, \bar{l}_K) = f_d(y|\bar{l}_K, d(\bar{l}_K))\prod_{j=0}^K f(l_j|\bar{l}_{j-1}, d(\bar{l}_{j-1}))$ since the terms $f_d(a_j|\bar{l}_j, \bar{a}_{j-1})$ are equal to 1 in expression (13) for $f_d(o)$. Thus the chance $f_d(y)$ of having $Y = y$ under $f_d(\cdot)$ is

$$\begin{aligned} f_d(y) &= \int f_d(y, \bar{l}_K) d\mu(\bar{l}_K) \\ &= \int \{f(y|\bar{l}_K, d(\bar{l}_K)) \times \prod_{j=0}^K f(l_j|\bar{l}_{j-1}, d(\bar{l}_{j-1}))\} d\mu(l_j) \end{aligned} \quad (14a)$$

where $d\mu(\bar{l}_K)$ represents integration (or summation if \bar{l}_K is discrete) over all possible \bar{l}_K histories. That is if \bar{l}_K is discrete we obtain

$$\begin{aligned} f_d(y) &= \sum_{\bar{l}_K} f_d(y, \bar{l}_K) \\ &= \sum_{\bar{l}_K} \{f(y|\bar{l}_K, d(\bar{l}_K)) \times \prod_{j=0}^K f(l_j|\bar{l}_{j-1}, d(\bar{l}_{j-1}))\} \end{aligned} \quad (14b)$$

with \bar{l}_j the initial segment of a given \bar{l}_K history. Robins (1986) referred to this formula as the g-computation algorithm formula or functional for the effect of regime d on outcome Y . Similarly, the marginal distribution function of Y under $F_d(\cdot)$ is

$$F_d(y) = \int \dots \int pr[Y < y|\bar{l}_K, d(\bar{l}_K)] \times \prod_{j=0}^K f(l_j|\bar{l}_{j-1}, d(\bar{l}_{j-1})) d\mu(l_j)$$

To estimate $f_d(y)$ from the observed data we specify parametric models $f(l_j|\bar{l}_{j-1}, \bar{a}_{j-1})$ and $f(y|\bar{l}_K, \bar{a}_K)$, fit the models to the observed data to obtain estimates $\hat{f}(l_j|\bar{l}_{j-1}, \bar{a}_{j-1})$ and $\hat{f}(y|\bar{l}_K, \bar{a}_K)$ and evaluate these estimates at $\bar{a}_{j-1} = d(\bar{l}_{j-1})$ and $\bar{a}_K = d(\bar{l}_K)$, and substitute them into the above expression for $f_d(y)$. Details are given in section 6 below where we consider estimating the effect of treatment on a survival outcome. Under certain additional assumptions described in section 6, one may consider the g-formula as giving the overall effect of an intervention regime d on Y by summing up both its indirect effects on Y mediated through the effect of the intervention on the non-intervened on variables L_j (which in turn affect the outcome Y) and its direct effects on Y . In contrast, MSMs always model the overall effect of the intervention on Y without any decomposition into direct and indirect effects.

6. ANALYSIS OF THE FRAMINGHAM OFFSPRING STUDY

In this section we describe the use of the parametric g-computation algorithm formula estimator of section 5 and IPTW MSM estimator of section 4 to estimate what the cumulative incidence of coronary heart disease mortality in the Framingham Offspring Study would have been under various hypothetical intervention strategies. The Framingham Offspring public use file contained data on 5124 subjects (2483 male and 2641 female subjects). Subjects were examined five times over 20 years. Exam 1 occurred in year 0, exam 2 in year 8, exam 3 in year 12, exam 4 in year 16 and exam 5 in year 20. The table below shows the number of subjects attending each of the first four exams. Mortality follow-up was essentially complete through exam 5 at year 20. By year 20, 891 subjects were known to have developed coronary heart disease (CHD), including those with pre-existing CHD at exam 1.

<i>Content</i>	<i>Time after follow-up</i>	<i>Sample size</i>
Exam 1	Year 0	5 124
Exam 2	Year 8	3 863
Exam 3	Year 12	3 873
Exam 4	Years 16	4 019

For reasons explained below, the follow-up period for our analysis began at exam 2 (which we will refer to as the baseline exam). Each subject contributed person-time from exam 2 to the date of CHD diagnosis, death from any cause, loss to follow-up, or exam 5, whichever occurred first. Subjects with pre-existing CHD at exam 2 or with missing risk factor values at exam 1 were excluded from all calculations.

In the Framingham public use file, CHD was defined as one of the following events: (i) recognized or unrecognized myocardial infarction (using diagnostic ECG or transaminase/history/autopsy evidence), (ii) angina pectoris (first episode), (iii) coronary insufficiency, or (iv) death from CHD. The following (time-dependent) risk factors were used in various analyses: age at exam (years); sex (male/female); body mass index (kg/m^2); oigarette smoking (current smoker/non-current smoker); alcohol consumption (calculated alcohol index in Ozs. per week); diabetes mellitus (defined diabetes/no defined diabetes); LDL-cholesterol (mg/dl); HDL-cholesterol (mg/dl); and systolic blood pressure (mmHg).

Missing values of risk factors in later exams were carried forward from the previous exam. Fat intake was not used in our analyses, because the Framingham public use file did not contain that information. Physical activity was also not used, because it was not measured at Exams 1 and 4.

6.1 INTERVENTIONS

We estimated with the parametric g-computation algorithm formula estimator the cumulative incidence (i.e. risk) of CHD between exam 2 and exam 5 that would have been observed under the following hypothetical intervention strategies.

1. A random half of smokers at baseline (i.e. exam 2) quit smoking forever (no intervention on those who were not current smokers at baseline but started smoking later).
2. All smokers at baseline quit smoking forever (no intervention on those who were not current smokers at baseline but started smoking later).
3. Subjects who were drinkers at baseline were randomly assigned a non-time-varying level of alcohol consumption at baseline drawn at random from a truncated normal distribution with a mean of 20 g/day for males and 10 g/day for females, a maximum of 30 g/day for males and 15 g/day for females, and a minimum of 10 g/day for males and 5 g/day for females (with the sex-specific variance of the normal equal to the sex-specific variance of alcohol consumption at baseline among drinkers in study population). Alcohol consumption was not intervened on thereafter.
4. Body mass index (BMI) was lowered
 - a) by 10% at each visit BMI exceeded 22
 - b) to 22 at each visit BMI exceeded 22
5. The distribution of LDL at baseline was drawn at random from a normal distribution with mean of 90 and a SD of 30 (Chinese distribution both for males and females). LDL was not intervened on thereafter.
6. Intervention 1, 3 and 4a simultaneously.

6.2 DATA ANALYSIS

PARAMETRIC G-FORMULA

In this section we describe how we estimate the proportion of the Framingham Offspring Study population that would have developed CHD between Exams 2 and 5 for each of the interventions mentioned above under the assumption of no unmeasured confounders. Specifically, we used the parametric g-formula (i.e. the g-formula evaluated at predicted values calculated under parametric regression models) to estimate the CHD risk in the Framingham Offspring data under each intervention. We remark that in this exercise we are estimating the cumulative incidence (probability of developing) CHD between Exams 2 and 5 under each intervention rather than the expected life years and expected

quality-adjusted life-years. Although these latter quantities are more relevant for public health decision-making, we decided to report the results in terms of cumulative incidences because this measure is more familiar to practising epidemiologists. It would be easy to use these same methods to estimate expected life measures. The g-formula to compute the CHD risk (i.e. cumulative incidences between Exam 2 and 5) under a particular intervention (i.e. a possibly dynamic regime) d is one minus the probability of “not developing CHD between Exams 2 and 5” which is given by the following g-formula for “survival without CHD.”

$$\begin{aligned} & \sum_{\bar{a}_4} \sum_{\bar{a}_j^*} \sum_{l_4} \prod_{j=1}^4 \Pr[\text{CHD}_{j+1} = 0 | \bar{l}_j, \bar{a}_j, \overline{\text{CHD}}_j = 0, \bar{D}_j = 0]^{I(j>2)} \\ & \quad \times f_d(a_j | \bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) \\ & \quad \times f(l_j | \bar{l}_{j-1}, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) \\ & \quad f(a_j^* | \bar{l}_j, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) \end{aligned} \quad (15)$$

where j denotes exam number

for any \bar{z}_j , \bar{z}_0 is defined to be 0 since there is no exam 0.

$\text{CHD}_{j+1} = 1$ if CHD was diagnosed between exams j and $j + 1$, 0 otherwise;

the overbar means history, i.e. $\overline{\text{CHD}}_j = (\text{CHD}_1, \dots, \text{CHD}_j)$;

the random variable L_j with realized values l_j is the set of risk factors not undergoing the intervention;

a_j is the intervention value of the risk factors A_j undergoing intervention at time j ;

\bar{l}_j is a history of (non-intervened on) risk factors through exam j compatible with a particular \bar{l}_4 history in the sum;

\bar{a}_j is a history of (intervened on) risk factors through exam j ;

a_j^* is the value of A_j that would be observed at time j if interventions were made through $j - 1$ but no intervention was made at j (see below);

the sum is overall possible $\bar{l}_4, \bar{a}_4, \bar{a}_4^*$ histories;

$\bar{C}_j = 0$ is the event that a subject remains uncensored through exam j ;

$\bar{D}_j = 0$ is the event that a subject has not died from other non-CHD causes through exam j ;

$f(\cdot | \cdot)$ is a conditional density function of the observed data distribution;

$f_d(\cdot | \cdot)$ is the conditional density function of the exposure A under the proposed intervention. Because we do not intervene until $j = 2$, the “intervened on” value a_1 of A_1 must equal the “non-intervened on” value a_1^* . To accomplish this, we take for $j = 1$ $f_d(a_j | \bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) = 1$ if $a_j = a_j^*$ and 0 otherwise.

Formula (15) includes four generalizations of formula (14). First we allow the intervention treatment a_j at time j to depend on the value a_j^*

of A_j that would be observed at time j if the planned interventions were made through $j - 1$ but no intervention was made at j . That is the intervention density $f_d(a_j|\bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ can be a function of a_j^* . This is meant to reflect the fact that when a subject arrives at visit j his exposure a_j^* could be noted and if it takes certain values it can, in principle, be intervened on and instantaneously changed to a new exposure labelled a_j . We assume it is a_j that affects outcomes at the next visit $j + 1$. For example, the BMI intervention 4b was modelled in this way. This “instantaneous change” model might be a reasonable approximation to a real intervention in which BMI was reduced to 22 over 3 to 6 months, as the time between visits is four years. To determine at time j the value of a_j^* for a simulated subject, we need to generate a_j^* from the non-intervention observed data density $f(a_j^*|\bar{l}_j, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$, where it follows from the definition of a_j^* that the arguments $(\bar{l}_j, \bar{a}_{j-1})$ in the conditioning event are the values that would be seen had the planned intervention d been made through $j - 1$. Second, we now allow probabilistic (i.e. random) interventions so $f_d(a_j|\bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ need not take only the values 0 or 1. For example, an intervention in which a subject who is a smoker at occasion j (i.e. $a_j^* = 1$) stops smoking with probability 1/2 would have $f_d(a_j|\bar{l}_j, a_j^* = 1, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) = 1/2$ for both $a_j = 1$ and $a_j = 0$. Third, we now allow the interventions such as the alcohol intervention 3 and LDL intervention 5 to include active intervention at some times and no further intervention at other times. For example, in intervention 5, $f_d(a_j|\bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ is equal to a normal density with mean 90 and standard deviation 30 at the baseline visit $j = 2$, but $f_d(a_j|\bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) = 1$ if $a_j = a_j^*$ at visits $j = 3$ and $j = 4$ in addition to visit $j = 1$. Finally, the end-point is now survival to visit 5, so the g-formula multiplies the probability of not failing at visits 3, 4 and 5.

Caveats

As discussed above, the g-formula identifies the effect one would have observed under a particular intervention under the assumption of no unmeasured confounders. But this assumption itself assumes that the counterfactual survival variables T_d (the time to CHD under intervention d) are well defined. If the intervention is on smoking, it seems only mildly philosophically problematic to assume the existence of a well-defined counterfactual time of CHD had one received a different smoking history than one's actual history. However, for interventions 4a, 4b or 5, the counterfactual T_d may be very poorly defined. For example, the effect on CHD of lowering LDL may possibly depend on the mechanism by which it is lowered. That is, the effect may be different depending on whether LDL were lowered by using drugs that decrease LDL production, drugs that increase its elimination or drugs that impair the absorption of fats from the gut. Similarly the effect on CHD of an intervention

to lower BMI may be different depending on the degree of calorie restriction imposed, the foods allowed in the required diet, and the level of increased physical activity imposed. Thus for the interventions 4a, 4b and 5d that are stated solely in terms of the level of BMI or LDL to be achieved (without further specifying the precise intervention that will be used to realize the reductions), it may be that the counterfactuals T_d should not be regarded as well-defined. But if the counterfactuals are not well-defined, the assumption of no unmeasured confounders is vacuous (as the assumption is in terms of the counterfactuals). As a consequence, for such interventions, it is unclear what the counterfactual quantity one hopes the g -formula to estimate. However, if one correctly held the belief that LDL in the blood is a dominant cause of CHD and that all reasonable interventions that lowered LDL to a prespecified level would result in equal reductions in CHD risk, then T_d would be well-defined even for an intervention d that, like intervention 5, is stated only in terms of the level of LDL to be achieved. Evidence for this belief would be strengthened if randomized experiments of various lipid lowering therapies that lowered average LDL by an identical amount showed the same improvement in CHD risk, regardless of the mechanism of action of the drug used to lower LDL.

Even when a counterfactual is well-defined, the g -formula only estimates the effect of an intervention under the assumption of no unmeasured confounders. Thus it is important for epidemiologists to try to obtain data on many potential confounders. The assumption of no unmeasured confounders will never exactly be true in observational studies and may not even be approximately true. A particular setting in which substantial unmeasured confounding may be present is when many of the confounders recorded in L_j are measured with error. In that case, even if the correctly measured confounders would have served to fully control confounding, the mismeasured confounders may not. This could be the case even when the measurement error is random and non-differential and the null hypothesis of no treatment effect is true. This raises an important distinction between studies with time-varying treatments and studies with time-independent treatments. With time-varying treatments, past treatment history itself can be a confounder for the effect of current treatment on the outcome. Thus even when treatment is only subject to random (non-differential) measurement error, confounding bias can exist even under the null hypothesis of no treatment effect. This is in sharp contrast with the results for time-independent treatments in which random (non-differential) subject-specific measurement error does not lead to bias under the causal null hypothesis of no treatment effect. However, there is an additional subtlety in studies with time-varying treatments. If the decision to take a treatment at time j depends on one's past measured treatment history (say as recorded in a pharmacy prescription database) and not on one's true past treatment history, then,

under the null hypothesis, there is no bias introduced by the mismeasurement of treatment, even if the measurement error is differential. On the other hand, if the decision to take treatment at a given time depends on past true (but unrecorded) treatment history, then, by only controlling for past measured treatment history in the analysis, bias may exist under the null hypothesis of no treatment effect, even with random (non-differential) measurement error.

In sections 4 and 5, we assumed the intervened on risk factors A_j at time or visit j occurred temporally after the non-intervened on risk factors L_j . We clearly cannot assume this to be the case in the Framingham Offspring Study for all 6 interventions of interest. Further, because the variables are generally measured but once every four years, each risk factor may potentially causally influence the others over the four years. In this setting, if we let Z_j be the set of all measured risk factors at visit j (regardless of whether one wants to consider intervening on any subset), let D_j be the indicator variable for survival at visit j with D_j prior to Z_j , then a sufficient and nearly necessary condition for the g-functional (15) to identify the effect on the cumulative incidence of CHD under an intervention on any chosen subset of the measured time-varying risk factors is that (i) no risk factor causally influences any other risk factor over the four years between visits which we formalize as, for each visit j , no element of Z_j is a cause of any other element of Z_j on the causal DAG generating the data, (ii) the set of all measured risk factors \bar{Z}_K are jointly causally exogenous for survival in the sense that on the causal DAG generating the data, for all combination of visit times s, t, j , there is no unmeasured U_s that is a common cause of both D_j and (an element of) Z_t for which there exist directed paths from U_s to D_j and U_s to Z_t whose other vertices are all unmeasured variables, and (iii) any unmeasured common causes of variables in Z_j are marginally independent of any unmeasured common cause of variables in Z_t for $j \neq s$. That is, conditions (i)–(iii) guarantee that there are no unmeasured confounders for the effect of any possible intervention on survival, even without knowledge of time order among the variables Z_t measured at the same visit t and even though there is substantial time (four years) between the visits at which data are recorded. Conditions (i)–(iii) are so restrictive that it seems unlikely that we would ever believe that the g-formula will be exactly unbiased for the effect of all the various risk factor interventions in which we might have interest. If conditions (i)–(iii) did hold, then the g-formula gives the overall effect of an intervention regime d on Y by explicitly summing up both its indirect effects on survival mediated through the effect of the intervention on the non-intervened on variables L_j measured by the factors $f(l_j|\bar{l}_{j-1}, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ (which in turn affect the survival as measured

by the dependence of the factors $\prod_{j=1}^4 \Pr[CHD_{j+1} = 0 | \bar{l}_j, \bar{a}_j, \overline{CHD}_j = 0, \bar{D}_j = 0]$

on \bar{l}_j) and its direct effects on survival measured by the dependence of

$$\prod_{j=1}^4 \Pr[\text{CHD}_{j+1} = 0 | \bar{l}_j, \bar{a}_j, \overline{\text{CHD}}_j = 0, \bar{D}_j = 0] \text{ on } \bar{a}_j.$$

Details of estimation

Because the g-formula is a sum over all possible values of risk factor history \bar{l}_4 and each l_j is a high-dimensional vector of covariates, a direct calculation based on (15) is computationally infeasible. Rather, we approximate the result of the g-formula under a given intervention by Monte Carlo simulation. To see how to conduct the simulation, first note that the g-formula (15) gives the probability of developing CHD between Exams 2 and 5 based on a intervention-specific joint distribution of CHD and risk factors. Under the assumption of no unmeasured confounders, this is the joint distribution had all subjects followed the intervention.

Therefore we generate, for each intervention, a simulated population in which the joint distribution of CHD and risk factors is approximately equal to the joint distribution implied by the g-formula. Then the CHD risk in the simulated population (i.e. the expected fraction of subjects in the simulated population who develop CHD between Exams 2 and 5) estimates the desired probability. Note that, if we simulate under no intervention (i.e. with $f_d(a_j | \bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) = 1$ if $a_j = a_j^*$ for all j rather than for just $j = 1$), the expected CHD risk in the simulated population should equal that of the actual study population, because the joint distribution implied by the g-formula for the simulated population is precisely that of the study population.

Let $z_j = (l_j, a_j^*)$. To estimate

$$\begin{aligned} & f(z_j | \bar{l}_{j-1}, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) \\ &= f(a_j^* | \bar{l}_j, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) \\ & f(l_j | \bar{l}_{j-1}, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0) \end{aligned}$$

where l_j includes all non-intervened on risk factors at exam j , we chose an arbitrary ordering of risk factors at exam j : such as (i) body mass index, (ii) cigarette smoking, (iii) alcohol consumption, (iv) diabetes mellitus, (v) LDL, (vi) HDL, and (vii) systolic blood pressure. The density (15) is invariant to the ordering of the variables in z_j . This invariance to ordering is one reason why (15) can only have a causal interpretation for all regimes d under that assumption that no risk factor measured at j causes any other risk measure at j , as in the above caveats. We then estimate (i) the conditional probability of BMI at j given the past variables through $j - 1$, (ii) the conditional probability of smoking given BMI at j and past variables through $j - 1$, (iii) the conditional probab-

ity of alcohol at j given BMI, smoking at j and past variables through $j - 1$, and so on. We estimate each of these conditional densities by maximum likelihood from the observed data. Finally, we estimate $f(z_j | \bar{l}_{j-1}, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ as the product of these estimated conditional densities.

In detail, the algorithm used to simulate the simulated population exposed to an intervention was as follows:

Part A: modelling

1. We fit a pooled (over persons and time) linear logistic regression model to predict the risk of CHD given risk factor history. The outcome was CHD diagnosis between exams j and $j + 1$, and the covariates in the model were risk factors at exams j and $j - 1$. The model was restricted to those with no diagnosis of CHD at or before exam j . The parameters of this model define the estimated conditional probability of CHD risk given the entire past, and thus implicitly assume that risk factors more than 2 time periods previously do not predict CHD risk given risk factors in the past 2 periods.
2. For $j = 3$ and $j = 4$, we fit pooled regression models to predict each risk factor given past risk factor history, among those with no prior diagnosis of CHD. Risk factors at exam j were the “outcome” in models that include other risk factors at exam j (according to the arbitrary ordering explained above) plus all risk factors at exams $j - 1$ and $j - 2$ as covariates. We used linear regression for continuous risk factors, and logistic regression for dichotomous risk factors. Continuous risk factor variables with skewed distributions were log transformed. The parameters of these models define the estimated conditional distributions (Bernoulli or Normal or Log normal) of each risk factor.
3. Follow-up started at exam 2 in our analyses because two prior exams are used to predict CHD risk between exams j and $j + 1$. Thus, the CHD risk we estimate refers to the 12-year period between exam 2 (year 8) and exam 5 (year 20).

Part B: Data generation

4. We simulated a sample of 10 000 individuals by sampling with replacement from the study population.
5. The risk factor values at exams 1 and 2 of these 10 000 individuals were those actually observed (as interventions beginning at exam 2 could not affect these distributions so we were able to use the empirical distribution of the data).
6. The risk factor values $z_j = (l_j, a_j^*)$ at exams 3 and 4 were generated by sampling a value from the conditional distributions estimated in step 2 above.



7. a_j was set equal to a_j^* for each j .
8. The 12-year probability $1 - \prod_{j=1}^4 \Pr[\text{CHD}_{j+1} = 0 | \bar{l}_j, \bar{a}_j, \overline{\text{CHD}}_j = 0, \bar{D}_j = 0]$ of CHD for each simulated individual was estimated, based on his/her simulated risk factor values, using the conditional distribution estimated in step 1 above. The population CHD risk is the average of each person's estimated risk.
9. The above procedure simulates the observed study population under no intervention.
10. To simulate a counterfactual population subject to a given intervention d under the assumption of no unmeasured confounders, step 7 had to be modified by drawing a_j for $j = 2, 3, 4$ from the intervention density $f_d(a_j | \bar{l}_j, a_j^*, \bar{a}_{j-1}, \overline{\text{CHD}}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ rather than setting a_j to a_j^* as under no intervention.

We used nonparametric bootstrap methods (sampling the observed study population with replacement 100 times) to estimate approximate confidence intervals (CI) of the counterfactual CHD risks and risk ratios. The size of each bootstrap sample was that of the original sample. To obtain confidence intervals, we re-applied all of steps 1–8 to each of the 100 bootstrap samples. All analyses were performed separately by sex.

MARGINAL STRUCTURAL MODEL

We also analysed the non-dynamic interventions on cigarette smoking using MSMs and compared the results to those obtained using the g-formula.

Specifically our goal was (i) to estimate the (potentially time-varying) hazard of CHD under pre-specified (i.e. non-dynamic) regimes or interventions, and (ii) to use this hazard to compute the 12-year CHD risk.

To estimate the hazard, we fit the discrete-time marginal structural Cox model

$$\log \text{it} \lambda_{T_{\bar{a}}}(t) = \log \text{it} \lambda_0(t) + \beta a(t)$$

where $\lambda_{T_{\bar{a}}}(t) = \Pr[\text{CHD}_{\bar{a}}(t+1) = 1 | \text{CHD}_{\bar{a}}(t) = 0]$ is the discrete hazard of CHD between t and $t+1$ under intervention $\bar{a} = [a(2), a(3), a(4)]$ at visits 2, 3, 4, $\lambda_0(t)$ is the baseline hazard, $a(t)$ is smoking status (smoker or non-smoker) at study visit t under regime \bar{a} , and, for dichotomous smoking exposure, β is the log relative risk (i.e. odds ratio) for always exposed vs never exposed. Under the assumption of no unmeasured confounders, the parameters of this model can be estimated using inverse-probability-of-treatment weighting. The model used to predict smoking status was the same as the one used for the parametric g-formula. We fit



the model by using a weighted pooled logistic model with a time-varying intercept.

To estimate the 12-year CHD risk when everybody quits smoking at baseline (and nobody initiates smoking thereafter), we could have computed one minus the survival probability

$$[1 - \lambda_{T_{\bar{a}}}(2)][1 - \lambda_{T_{\bar{a}}}(3)][1 - \lambda_{T_{\bar{a}}}(4)]$$

with $\bar{a} = [0, 0, 0]$. However, this simple way of estimating the risk is potentially very inefficient (large variance) because the inverse-probability-of-treatment weights can be unstable. Therefore, to improve efficiency we used inverse-probability-of-treatment weighting to estimate the parameters of the model

$$\text{logit}\lambda_{T_{\bar{a}}}(t) = \text{logit}\lambda_0(t) + \alpha^T V + \beta a(t)$$

where V is a vector of baseline (pre-intervention) covariates. The inclusion of V allows us to use V -stabilized weights which result in much more efficient estimation.

We then simulated a sample of 10 000 individuals by sampling with replacement from the study population and used the parameter estimates of the MSM above to compute the CHD risk between Exams 2 and 5 for each simulated individual at each time based on the observed values of V and $\bar{A}(t)$. We then computed the 12-year CHD risk for each individual and averaged over all individuals to obtain an estimate of the risk under no intervention.

To estimate the 12-year CHD risk when 50% of the subjects quits smoking at baseline (and nobody initiates or quits smoking thereafter), we first assigned a random half of the smokers in our simulated sample to quitting smoking and then proceeded as above to compute each subject's 12-year CHD risk under his/her smoking history (always smoking or never smoking).

7. RESULTS

After exclusions, our analyses included 2230 men (47.8%) and 2440 (52.2%) women with 189 and 68 CHD events, respectively. Thus, the observed 12-year risk of CHD in the study population was 8.48% (95% CI 7.37%–9.73%) for males and 2.79% (95% CI 2.19%–3.54%) for females.

The estimated 12-year risks of CHD (and 95% CI) based on the parametric g -formula under several public health interventions are shown in Table 28.1. The simulated 12-year risk of CHD under no intervention was 8.46% for males (95% CI 5.61%–11.30%) and 2.82% for females (95% CI 2.15%–3.49%). The risk ratios (and 95% CIs) for each intervention compared with no intervention are also shown in Table 28.1.

For example, the estimated risk ratio among men is 0.81 (95% CI 0.73–0.90) for all smokers quitting smoking at baseline vs no intervention. Note in row 6, we obtain the overall effect of combined interventions without explicitly imposing any particular functional form (e.g. multiplicative or additive) for the interaction of the individual interventions 1, 3 and 4a on the overall risk of CHD.

Table 28.2 displays the results when only one repeated measure of each risk factor (the most recent one) was included in the model for CHD risk. Results in Table 28.1 come from a model that includes the two most recent measures of each factor. The results in the two tables are compared and contrasted in the next section. Results of MSM analyses for the smoking intervention are given in Table 28.3. Table 28.3 displays the risk and risk ratio estimates from a MSM for two smoking interventions similar to interventions 1 and 2, respectively.

8. DISCUSSION

We have presented two methods—parametric g-formula and MSMs—to estimate the causal effect of hypothetical public health interventions in the Framingham Offspring Study. In the absence of unmeasured confounders and model misspecification, these methods provide causal estimates from observational data that mimic the results of randomized experiments.

The dataset we used has two major limitations: (i) not all relevant confounders are available and (ii) the number of events, especially for women, is small. Because our methods are only valid when the assumption of no unmeasured confounders holds, we would not expect our estimates to be necessarily close to those of a randomized experiment. However, our methods gave coherent results, i.e. when applied to similar interventions, both methods yielded similar estimates. Such coherent results are some evidence against model misspecification but constitute no evidence against confounding by unmeasured factors as even when such unmeasured confounders exist both the parametric g-formula and MSM approaches are estimating the same association parameter. We now discuss the relative advantages and disadvantages of the parametric g-formula and MSMs.

8.1 MODEL MISSPECIFICATION

When using the parametric g-formula, gross model misspecification can be detected by comparing the observed risk and the estimated risk under no intervention. In our example, both risks are similar (8.48% vs 8.46% in men, 2.79% vs 2.82% in women). Note that, whereas dissimilar risks indicate model misspecification, similar risks cannot rule out the existence of model misspecification.

Even in the absence of unmeasured confounding, (i) valid estimation of the effect of the smoking intervention (1) with the parametric g-

Table 28.1 G-formula estimates (reference analysis)

<i>Intervention</i>	<i>Risk (%)</i>	<i>95% CI</i>	<i>Risk Ratio</i>	<i>95% CI</i>	<i>BS SE</i>	<i>BS av. risk</i>
Male, n = 2230						
0) No intervention (simulated results)	8.46	5.61–11.30	1.00		1.45	8.72
1) 50% quit smoking at baseline	7.65	5.02–10.27	0.90	0.86–0.95	1.34	7.86
2) All quit smoking at baseline	6.82	4.35–9.29	0.81	0.73–0.90	1.26	7.00
3) Alcohol intake to specified distribution	8.41	5.52–11.30	0.99	0.97–1.02	1.47	8.70
4a) BMI lowered by 10% when BMI >22	8.09	4.57–11.61	0.96	0.74–1.24	1.80	8.39
4b) BMI lowered to 22 when BMI >22	7.78	3.84–11.72	0.92	0.65–1.30	2.01	8.25
5) LDL shifts to Chinese distribution	5.80	3.35–8.25	0.69	0.57–0.82	1.25	6.17
6) Combined interventions 1, 3, and 4a	7.26	4.00–10.53	0.86	0.66–1.12	1.67	7.54
Female, n = 2440						
0) No intervention (simulated results)	2.82	2.15–3.49	1.00		0.34	2.79
1) 50% quit smoking at baseline	2.63	1.96–3.29	0.93	0.84–1.04	0.34	2.62
2) All quit smoking at baseline	2.44	1.65–3.22	0.86	0.68–1.09	0.40	2.46
3) Alcohol intake to specified distribution	2.83	2.15–3.51	1.00	0.95–1.06	0.35	2.81
4a) BMI lowered by 10% when BMI >22	3.11	2.20–4.02	1.10	0.90–1.34	0.46	3.06
4b) BMI lowered to 22 when BMI >22	3.23	1.94–4.51	1.14	0.82–1.59	0.66	3.23
5) LDL shifts to Chinese distribution	1.44	0.72–2.15	0.51	0.34–0.77	0.37	1.51
6) Combined interventions 1, 3, and 4a	2.91	1.95–3.87	1.03	0.80–1.33	0.49	2.91

Key: BS SE, bootstrap standard error; BS av. risk, bootstrap average risk.

formula requires that one can correctly model both (a) the conditional probability of developing CHD conditional on risk factor history and (b) the conditional probability of current non-intervened on risk factors given past risk factor history; while (ii) valid estimation with MSMs requires that one (a) correctly model the conditional probability of the intervened on risk factor (smoking) given risk factor history and (b) the structural discrete time proportional hazard model. The parametric g-formula does not require that one can correctly model the conditional probability of the intervened-on risk factor (smoking) given risk factor history because for the smoking intervention (1) the density $f_d(a_j^*|\bar{I}_j, a_j^*, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ does not actually depend on a_j^* so no model for $f(a_j^*|\bar{I}_j, a_j^*, \bar{a}_{j-1}, \overline{CHD}_j = 0, \bar{C}_j = 0, \bar{D}_j = 0)$ is required in (15) because the a_j^* simply get summed out. Thus the models required for valid estimation of the effect of the smoking intervention (1) with the parametric g-formula have no overlap with the models required for valid estimation with MSMs. As a result, if estimates of the effect of the intervention agree under the two approaches, one can be somewhat confident that major model misspecification is absent. This represents a great

Table 28.2 G-formula estimates (risk factor history lagged 1 visit only)

<i>Intervention</i>	<i>Risk (%)</i>	<i>95% CI</i>	<i>Risk Ratio</i>	<i>95% CI</i>	<i>BS SE</i>	<i>BS av. risk</i>
Male, n = 2 230						
0) No intervention (simulated results)	8.46	5.55–11.36	1.00		1.48	8.72
1) 50% quit smoking at baseline	7.63	4.92–10.34	0.90	0.86–0.95	1.38	7.85
2) All quit smoking at baseline	6.79	4.20–9.37	0.80	0.72–0.90	1.32	6.98
3) Alcohol intake to specified distribution	8.42	5.48–11.36	1.00	0.97–1.02	1.50	8.69
4a) BMI lowered by 10% when BMI > 22	7.74	4.49–10.99	0.92	0.78–1.08	1.66	8.06
4b) BMI lowered to 22 when BMI > 22	7.40	3.86–10.94	0.88	0.69–1.11	1.80	7.75
5) LDL shifts to Chinese distribution	5.80	3.24–8.36	0.69	0.57–0.82	1.31	6.11
6) Combined interventions 1, 3, and 4a	6.94	3.87–10.01	0.82	0.69–0.98	1.57	7.22
Female, n = 2 440						
0) No intervention (simulated results)	2.80	2.15–3.45	1.00		0.33	2.74
1) 50% quit smoking at baseline	2.57	1.94–3.21	0.92	0.83–1.02	0.32	2.54
2) All quit smoking at baseline	2.36	1.64–3.08	0.84	0.68–1.05	0.37	2.34
3) Alcohol intake to specified distribution	2.79	2.14–3.43	1.00	0.96–1.03	0.33	2.72
4a) BMI lowered by 10% when BMI >22	2.76	2.05–3.48	0.99	0.83–1.18	0.36	2.69
4b) BMI lowered to 22 when BMI >22	2.73	1.91–3.55	0.98	0.76–1.26	0.42	2.65
5) LDL shifts to Chinese distribution	1.36	0.82–1.91	0.49	0.35–0.68	0.28	1.36
6) Combined interventions 1, 3, and 4a	2.53	1.82–3.24	0.90	0.73–1.12	0.36	2.48

Key: BS SE, bootstrap standard error; BS av. risk, bootstrap average risk.

Table 28.3 Estimates from marginal structural model

<i>Intervention</i>	<i>Risk (%)</i>	<i>95% CI</i>	<i>Risk Ratio</i>	<i>95% CI</i>	<i>BS SE</i>	<i>BS av. risk</i>
Male, n = 2 230						
0) No intervention (simulated results)	8.61	6.97–10.24	1.00	0.83	8.55	
1) 50% quit smoking at baseline	7.93	6.61–9.25	0.92	0.85–1.00	0.67	7.86
2) All quit smoking at baseline	6.93	5.52–8.34	0.80	0.65–1.00	0.72	6.89
Female, n = 2 440						
0) No intervention (simulated results)	2.57	1.96–3.19	1.00		0.32	2.41
1) 50% quit smoking at baseline	2.37	1.78–2.96	0.92	0.78–1.09	0.30	2.26
2) All quit smoking at baseline	2.10	1.29–2.91	0.82	0.55–1.22	0.42	2.06

Key: BS SE, bootstrap standard error; BS av. risk, bootstrap average risk.

advantage of having two independent methods of estimating the same intervention effect; each method can either reinforce or call into question the results obtained under the alternative method.

A subtle form of model misspecification when using the parametric g-formula is due to collinearity among the repeated measures of the risk factors. Collinearity is not a problem for the prediction of CHD risk under no intervention but it may be a problem for predictions of CHD risk under certain interventions. For example, suppose that

1. in the logistic model for CHD risk at time 3, the odds ratio for log *BMI* at time 2 is 0.2 and the odds ratio for log *BMI* at time 1 is 5, and because of collinearity among the repeated measures of BMI, neither estimate is significantly different from zero, but a chi-squared test on two degrees of freedom of the hypothesis that both coefficients are zero strongly rejects.
2. In the observed data, a subject has a quite elevated BMI of 29 at Exams 1 and 2. Nonetheless, the predicted effect of the subjects BMI at Exams 1 and 2 on CHD risk at Exam 3 is null in the sense that the odds ratio comparing any level of BMI that is the same at the two exams with any other level of BMI also constant at the two Exams on CHD risk at Exam 3 is $0.2 \times 5 = 1.0$.

Under 1 and 2, the prediction of CHD risk for the subject under no intervention is quite stable (i.e. the confidence interval for the predicted probability is narrow). Now suppose we intervene to set BMI equal to 22 as soon as it is greater than 22 for times greater or equal than 2 (intervention 4b). Then the overall effect of BMI is predicted to be so harmful that we are practically killing the subject since the odds ratio is $5^{\ln(29-22)} = 22.9$ compared to a subject who has BMI of 29 at both exams (or to a subject who has a BMI of 22 at both exams). However, in any given bootstrap sample, the odds ratios may be reversed due to sampling variability (because of the high correlation between BMI values over time). Therefore, in some bootstrap samples we would appear to be effectively killing the subject, whereas in others BMI would appear to have a strong protective effect. The net result is the large variance we found for the risk (and risk ratio) of intervention 4b and, to a lesser extent, of intervention 4a (Table 28.1). The variance of the risk estimate for intervention 4b was 39% greater than that of the risk estimate under no intervention.

To further assess issues of model misspecification, we conducted two separate sensitivity analyses: (i) we added a quadratic term to the linear term for log BMI in the CHD model, and (ii) we used cubic splines with 3 knots for log BMI in the CHD model. Both of these strategies slightly increased the variance of the estimates for intervention 4a or 4b (data not shown) and the risk ratio estimates did not change significantly.

The high variability arises because the correlation structure of the data is destroyed under the intervention. That is, there is no subject in the sample who has a BMI of 29 at occasion 1 and of 22 at occasion 2. Thus, it is not surprising and wholly appropriate that we are uncertain of what the result would be of an intervention that would create such subjects. Indeed, because there is no subject in the sample who has a BMI of 29 at occasion 1 and of 22 at occasion 2, the uncertainty we see in Table 28.1 for intervention 4b is really an underestimate; it would have been more appropriate to admit that there was no data evidence with which to estimate the effect of such an intervention. Our apparent ability to estimate the intervention effect at all (albeit with great uncertainty) was wholly based on extrapolation under the, possibly incorrect, modelling assumption that the effect of BMI at the last 2 visits on CHD risk is linear on a logistic scale.

Some investigators, when faced with the problem of highly collinear repeated measures of exposures and the associated high variance of predicted interventions, decide to enter only one of the measures in their model, incorrectly arguing that the collinearity obscures the true exposure effect. To empirically examine some of the consequences of this flawed logic, we included only the most recent measure of each risk factor in the CHD model in Table 28.2 rather than the most recent two measures as in Table 28.1. The variance of the risk estimate for intervention 4b was now only 24% greater than that of the risk estimate under no intervention. The difficulty, however, is that this estimate of the intervention effect can be badly biased if, in fact, the second most recent measure of BMI is in truth a risk factor for CHD controlling for the most recent BMI measure.

The lesson to be learned from the above discussion is that interventions that propose big and abrupt changes in the value of a risk factor that otherwise shows little variation over time (under no intervention) will yield causal estimates with large variances and even these large variances will underestimate the actual uncertainty. This will be true regardless of the methodology used (i.e. MSMs or the parametric g-formula). Intervention involving these risk factors should be formulated in ways that do not imply big and abrupt changes. For example, reducing BMI by 10% (intervention 4a) resulted in a smaller variance than reducing BMI to the value 22 (intervention 4b).

8.2 EFFICIENCY

The parametric g-formula yielded narrower confidence intervals for the risk ratios of interest than the MSM. This was expected based on underlying statistical theory.

8.3 TYPES OF INTERVENTIONS

Marginal structural models are not useful to estimate risks under dynamic interventions (i.e. interventions that depend on the evolving

values of time varying risk factors), such as interventions 4a and 4b. In those cases, the g -formula or structural nested models are needed.

A related subtle limitation of MSMs is that they cannot estimate the risk under no intervention in the presence of effect modification by time dependent non-intervened-on risk factors. This is so because IPTW estimation of a MSM effectively creates a pseudo-population in which the probability of receiving exposure does not depend on one's history of time dependent non-intervened-on risk factors. To clarify this point, we consider the following two scenarios:

1. Exposure is assigned to those who actually received it in the observed data
2. Exposure is assigned to random subjects (in the same proportion as in the original population)

The proportion of CHD in both scenarios will be equal only in the absence of effect modification. A MSM provides estimates of CHD risk for scenario 2 but not for scenario 1. One consequence of this limitation is that, when using MSMs, we lack a way to detect gross model misspecification by comparing observed and estimated risk under no intervention. In our example, the observed and estimated risks under no intervention were close, so we assumed that effect modification was not an important issue and the relative risks in Table 28.3 are good approximations. Note that the no-intervention MSM results represent scenario 2 while the empirical CHD risk in the cohort represents scenario 1.

Because of these same reasons, we needed to slightly modify the meaning of interventions 1 and 2 when using a MSM. Rather than allowing non smokers to follow their own observed history after baseline (as was done in Tables 28.1 and 28.2), we forced them to remain as non-smokers. This made little substantive difference as very few subjects initiated smoking after baseline.

8.4 CENSORING

We defined as censored all subjects that were either lost to follow-up before the last study visit or died from competing causes (i.e. not CHD). Our causal estimates can be interpreted as the CHD risk under each intervention had censoring been abolished (i.e. had nobody been lost to follow-up or died from competing causes), provided there is no unmeasured confounding and censoring is ignorable in the sense that the conditional cause-specific hazard of being censored at k among subjects alive and uncensored up to k does not depend on the time T to CHD given both $\bar{A}(k-1)$ and the time-dependent covariate $\bar{L}(k-1)$ history prior to k . However, one might be interested in the CHD risk had censoring by loss to follow-up, but not by competing causes of death, been abolished, especially because one may feel that the time T to CHD were

deaths from competing causes abolished is not a well-defined counterfactual quantity. For example, imagine a cohort of 100 people with 10 CHD cases, 20 deaths from other causes, and 70 subjects alive at the end of follow-up. If censoring from other causes was completely independent of failure and censoring by loss to follow-up, our CHD risk estimate would be $10/80 = 12.5\%$ whereas the risk estimate had other causes of death not been abolished is $10/100 = 10\%$. Of course, these two risks would only differ substantially if a large proportion of people died from competing causes, which is not the case in our data. We could easily adapt our methods to estimate the effect of interventions when deaths from other causes are included rather than eliminated.

In summary, estimation of non-dynamic interventions should be done both with MSMs and with the parametric g-formula, as agreement between the results is important evidence of lack of major model misspecification. MSMs are only useful to estimate the effect of non-dynamic interventions. The g-formula and structural nested models should be used to estimate the causal effect of dynamic interventions. In future work, we plan to compare g-formula and structural nested models based estimates of the effect of dynamic interventions in a manner similar to the comparisons of g-formula- and MSM-based estimates of non-dynamic regimes reported here.

9. CONCLUSIONS

We have tried to show why it is that even with the high quality longitudinal data collected in the Framingham Offspring Study, valid estimation from observational data of the effect of considered interventions requires one to make a large number of unverifiable assumptions that will never hold exactly and may not even hold approximately. In most of the world and for many interventions of interest, relevant high quality longitudinal data are unavailable, compounding the difficulties. Even if we succeeded in validly estimating the effect of simultaneous interventions on smoking, LDL, BMI and alcohol among the subjects in the Framingham Offspring Study, the question of how to extrapolate such results to other populations would remain unresolved. In this chapter we have stressed the large number of untestable assumptions that must be fulfilled to obtain valid estimates of causal effects and the improbability that all of them hold. We have not done so to discourage attempts to prioritize among potential interventions. Such prioritizing is and will continue to be done implicitly or explicitly every day. Our goal was simply to try to help the process by highlighting some of the central issues and potential biases that need to be carefully considered. It is a fact of life that deciding among potential interventions must be done—and must be done under great uncertainty. Methods for making decisions under uncertainty are well established, but the usefulness of such methods

ultimately rests on an honest and comprehensive assessment of the uncertainties. It is our hope that this chapter will help in making such assessments.

REFERENCES

- Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48.
- Hernán MA, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11:561–570.
- Hernán MA, Brumback B, Robins JM (2001) Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association—Applications and Case Studies*, 96: 440–448.
- Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47:663–685.
- Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.
- Manton KG, Stallard E, Singer B (1992) Projecting the future size and health status of the US elderly population. *International Journal of Forecasting*, 8:433–458.
- Neyman J (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5:465–480.
- Pearl J (1995) Causal diagrams for experimental research. *Biometrika*, 82: 669–710.
- Robins JM (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, with 1987 Errata in *Computers and Mathematics with Applications*, 14:917–921.
- Robins JM (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, with 1987 Errata in *Computers and Mathematics with Applications*, 18:477.
- Robins JM (1999) Marginal structural models versus structural nested models as tools for causal inference. *Statistical models in epidemiology: the environment and clinical trials*. IMA Volume 116. Halloran ME, Berry D, eds. Springer-Verlag, New York.
- Robins JM, Blevins D, Ritter G, Wulfsohn M (1992) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, 3:319–336. Please see also the following link for the Errata to the article: Robins JM, Blevins D, Ritter G, Wulfsohn M (1993) Errata to G-estimation of the effect of prophylaxis therapy for pneu-

- mycobacterium tuberculosis pneumonia on the survival of AIDS patients. *Epidemiology*, **4**:189. Reproduced with permission of the publisher, Lippincott, Williams and Wilkins.
- Robins JM, Greenland S, Hu F-C (1999) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, **94**:687–700.
- Robins JM, Rotnitzky A, Scharfstein D (1999) Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: *Statistical models in epidemiology: the environment and clinical trials*. Halloran ME, Berry D, eds. IMA Volume 116. Springer-Verlag, New York.
- Rubin D (1978) Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, **6**:34–58.
- Spirtes P, Glymour C, Scheines R, eds (1993) *Causation, prediction and search*. Springer-Verlag, New York.