

## ORIGINAL ARTICLE

# Causal models for estimating the effects of weight gain on mortality

JM Robins<sup>1,2</sup>

<sup>1</sup>Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA and <sup>2</sup>Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Suppose, in contrast to the fact, in 1950, we had put the cohort of 18-year-old non-smoking American men on a stringent mandatory diet that guaranteed that no one would ever weigh more than their baseline weight established at the age of 18 years. How would the counterfactual mortality of these 18 year olds have compared to their actual observed mortality through 2007? We describe in detail how this counterfactual contrast could be estimated from longitudinal epidemiologic data similar to that stored in the electronic medical records of a large health maintenance organization (HMO) by applying g-estimation to a novel of structural nested model (SNM). Our analytic approach differs from any alternative approach in that, in the absence of model misspecification, it can successfully adjust for (i) measured time-varying confounders such as exercise, hypertension and diabetes that are simultaneously intermediate variables on the causal pathway from weight gain to death and determinants of future weight gain, (ii) unmeasured confounding by undiagnosed preclinical disease (that is, reverse causation) that can cause both poor weight gain and premature mortality (provided an upper bound can be specified for the maximum length of time a subject may suffer from a subclinical illness severe enough to affect his weight without the illness becomes clinically manifest) and (iii) the presence of particular identifiable subgroups, such as those suffering from serious renal, liver, pulmonary and/or cardiac disease, in whom confounding by unmeasured prognostic factors is so severe as to render useless any attempt at direct analytic adjustment. However, (ii) and (iii) limit the ability to empirically test whether the SNM is misspecified. The other two g-methods—the parametric g-computation algorithm and inverse probability of treatment weighted estimation of marginal structural models—can adjust for potential bias due to (i) but not due to (ii) or (iii).

*International Journal of Obesity* (2008) 32, S15–S41; doi:10.1038/ijo.2008.83

**Keywords:** BMI; confounders; g-estimation; reverse causation; structural nested failure time model

## Introduction

Suppose, in contrast to the fact, in 1950, we had put the cohort of 18-year-old non-smoking American men on a stringent mandatory diet that guaranteed that no one would ever weigh more than their baseline weight established at the age of 18 years. Specifically, each subject was weighed every day starting on the day before his eighteenth birthday. Whenever his weight was greater than or equal to this baseline weight, the subject's caloric intake was restricted, without changing his usual mix of calorie sources and micronutrients, until the time (usually within 1–3 days) that the subject fell below baseline weight. (I restrict to men solely to avoid the complicating issue of how much weight gain to allow during pregnancy.) Thus, ignoring errors of a

pound or two, no subject would ever weigh more than his baseline weight. No instructions or restrictions were given concerning exercise at any time or the amount or nature of what the subject ate during non-calorie-restricted periods. How would the counterfactual mortality of these 18 year olds have compared to the actual observed mortality through 2007?

Factually, a substantial fraction of 18-year-old American male gains more than 30 pounds from age 18 to 74 years. Thus, if the counterfactual mortality were much less than the observed mortality, then, it would make sense for individuals to maintain their baseline body weight by restricting caloric intake (regardless of whether or not a practical, non-mandatory public health intervention exists that would successfully maintain the baseline weight of most of the (non-smoking) US population). Here and throughout, we use the phrase 'maintain their age  $x$  body-weight' to mean that after age  $x$  a subject's weight never exceeds his weight at the age of  $x$ , although it may drop below that weight.

Correspondence: Dr JM Robins, Department of Epidemiology, and Department of Biostatistics, Harvard School of Public Health, Kresge Building, Room 821, 677 Huntington Avenue, Boston, MA 02115, USA.  
E-mail: robins@hsph.harvard.edu

The difference between the counterfactual mortality where no one exceeds their age 18 body weight and the actual observed mortality of the non-smoking US population has been discussed by Willett *et al.*<sup>1</sup> as a useful way to conceptualize the effect of weight on mortality. A major goal of this paper is to show that g-estimation of structural nested models (SNMs) can be used to directly estimate this difference from longitudinal observational data. An SNM is a model that takes as inputs a subject's observed outcome, observed exposure (here, weight) history, and an unknown parameter and outputs the response that would have been observed if, possibly contrary to fact, the subject followed the stringent mandatory diet described above. The unknown parameter vector of an SNM is estimated through the g-estimation procedure introduced in Robins.<sup>2</sup> Previous analytic approaches to the estimation of the effect of weight on mortality do not provide a direct estimate of this difference. In addition, previous approaches have suffered from one or more of the following sources of bias: (i) failure to adequately control for measured confounding due to time-varying exercise, blood lipids, blood pressure (BP), diabetes (Db) and other chronic diseases (once diagnosed) because of concerns that one will thereby be controlling for intermediate variables on the causal pathway from overweight to death, (ii) failure to adequately control for unmeasured confounding due to undiagnosed chronic disease such as cancer (that is, reverse causation) and (iii) failure to update the weight of a subject whose weight changes after start of follow-up, because of concerns about reverse causation and measurement error.

Bias due to confounding by measured time-varying confounders that are also intermediate variables can be controlled by the use of the so-called g-methods. g-Methods are statistical methods specifically designed to control bias attributable to time-varying confounders affected by previous exposure. In addition to g-estimation of SNMs, g-methods include the parametric g-formula estimator and inverse probability of treatment weighted (IPTW) estimators.<sup>3,4</sup> As yet g-methods have not been used to estimate the effect of overweight on obesity with the exception of Robins *et al.*,<sup>5</sup> where the parametric g-formula estimator was used. In this paper, we concentrate on g-estimation of SNMs, because as discussed below, of the three g-methods, only g-estimation of SNMs can adjust for unmeasured confounding due to undiagnosed chronic disease.

Finally, g-estimation of SNMs allows one to update the weight of a subject whose weight changes after start of follow-up without introducing any bias due to reverse causation. However, issues of measurement error are trickier and will be discussed in the final section of the paper.

Even if maintenance of age 18 weight improves mortality, perhaps a mandatory intervention that allowed weight gain of 0.3/12 pounds per month (that is, 3 pounds per decade) would produce an even lower mortality. Perhaps the mandatory intervention that would produce the lowest mortality (that is, the optimal intervention among all

'weight-gain' interventions<sup>4,10</sup>) is one that allows a weight gain of 0.3/12 pounds per month in subjects free of hypertension, Db, hyperlipidemia or clinical CHD, but of only 0.1/12 pounds per month (that is, 1 pound per decade) once a subject developed one of these risk factors.

To decide which mandatory intervention is optimal, we require a well-defined numerical measure of overall mortality that can be used to rank interventions. For example, one might use the total years of life (or quality-adjusted life) experienced by the cohort from 1950 to 2007 as a measure. Use of this measure is mathematically equivalent to the use of 'years (or quality-adjusted years) of life lived from 1950 to 2007' as the (subject specific) utility function in a decision problem whose goal is to maximize expected utility. 'Years (or quality-adjusted years) of life lived' measures have a much more natural and useful public health and policy interpretation than the rate ratio, attributable fraction and attributable risk measures routinely reported in epidemiologic studies.

However, even 'years of life lived from 1950 to 2007' is an inadequate utility function when follow-up of the cohort is not to extinction. This function inappropriately assigns the same utility not only to all subjects alive at the age of 74 years on 1 January 2008 regardless of their state of health but also to a subject who dies on 31 December 2007 at 23:59 hours. Clearly among survivors in 2008, the healthier ones (according to some agreed standard measure of current health) have a greater post-study expected survival (and thus warrant a higher utility) than the less healthy survivors and a much greater expected survival (and thus warrant a much higher utility) than the non-survivors who died in late December 2007. We will not discuss further precisely how to decide on an appropriate utility measure for the survivors, except to remark that such a discussion is necessary. Rather, we will simply assume that, at the end of follow-up, each cohort member has been given a utility measure  $Y$ .

Note that the benefit of any of the above counterfactual interventions is an overall effect of the intervention. For example it is conceivable that the mortality benefit of the intervention that maintained baseline weight was wholly due to changes in exercise. Perhaps maintenance of baseline weight makes individuals feel so much better that they exercise more.

In the section Estimation of an overall effect, I assume we have observational retrospective follow-up data through 2007 on a random sample of the cohort of US males who were non-smokers and 18 in 1950. The data include detailed medical records, analogous to those currently available on subscribers to a comprehensive health maintenance organization (HMO). In the section Potential for measured and unmeasured confounding, I discuss three major sources of potential bias that complicate any attempt to estimate the overall effect of the mandatory intervention 'maintain baseline weight' on the expected utility of our cohort: (i) measured time-varying confounders such as exercise, hypertension and Db that are potentially intermediate variables, (ii) unmeasured confounding by undiagnosed preclinical disease (that is, reverse causation) that can cause both poor

weight gain and premature mortality and (iii) the presence of particular identifiable subgroups, such as those suffering from serious renal, liver, pulmonary and/or cardiac disease, in whom confounding by unmeasured prognostic factors is so severe as to render useless direct analytic adjustment for confounding. In the section Estimation of the effect of the ‘maintain baseline weight intervention,’ I describe how g-estimation of a correctly specified SNM can appropriately adjust for these potential sources of bias (provided an upper bound can be specified for the maximum length of time a subject may suffer from a subclinical illness severe enough to affect his weight before the illness becomes clinically manifest). The SNM required for this adjustment is novel in two ways. First it is a joint SNM, combining a structural nested failure time model (SNFTM) for the counterfactual time to the earlier of death or the diagnosis of a chronic illness and a conditional structural nested mean model (SNMM) for the mean of a subject’s counterfactual utility given his counterfactual time to death or a diagnosed chronic illness. Second our SNM only models the causal effect of any increase in BMI between month  $m$  and  $m+1$  over a subject’s maximum previous BMI. In particular, it does not model and thus is agnostic about the causal effect a) of any decrease in BMI or b) of any increase in BMI between  $m$  and  $m+1$  that fails to attain the previous maximum. As a consequence, our SNM is more robust than standard SNMs that also model a) and b), because our model makes fewer assumptions than such alternative models, and thus is less likely to be misspecified. However, the small number of assumptions made by our SNM are sufficient to consistently estimate the parameter of interest. In the sections Are remarkable results due to some sleight of hand and Can we replace  $X_m$  by  $X$  revisited, however, I show that (ii) and (iii) limit the ability to empirically test whether the joint SNM is misspecified. I also show that, somewhat remarkably, to adjust for bias due to reverse causation one need not assume a deterministic rank-preserving (RP) SNM. This is important as a deterministic RP SNM assumes that the effect of weight gain on mortality is the same for all subjects, an assumption that is clearly biologically implausible. In the section Censoring, I consider how to account for censoring by administrative end of follow-up. In the section Maximum weight gain dietary intervention regimens, I consider the estimation of the expected utility under alternative dietary interventions. In the section Measurement error, I discuss the consequences of measurement error in body mass index (BMI). Proofs and statements of several new theorems are collected in Appendices 1 and 2. Finally, estimation of the optimal ‘weight gain’ intervention is discussed in Appendix 3.

## Estimation of an overall effect

### The data

I describe the observational data that is supposed to be available. First, I suppose that a subject’s BMI is recorded at

the end of each month  $t$ ,  $t=0, 1, \dots, K$ , where *time*  $t$  is in months since age 18 years and  $K+1 = (2007-1950) \times 12$  is the duration of follow-up. Let  $A^*(t)$  be the difference between BMI at the end of month  $t$  and at the end of month  $t-1$ . Let  $L(t)$  be the vector of covariate values recorded in month  $t$  and suppose  $L(t)$  precedes  $A^*(t)$  temporally.  $L(t)$  includes BP, high-density lipoprotein and low-density lipoprotein (LDL) measures of cholesterol, any diagnoses of and clinical and laboratory characteristics of any chronic disease such as cancer, coronary artery disease, Db, asthma, chronic obstructive pulmonary disease, liver, renal disease, etc., level of exercise, measures of mobility and disability, etc. The vector  $L(t)$  also includes  $BMI(t)$ , the BMI just before the beginning of month  $t$  rounded to the nearest pound. Thus,

$$A^*(t) = BMI(t+1) - BMI(t) \quad (1)$$

$L(t)$  also includes the indicator  $I(T>t)$  of vital status at the beginning of month  $t$  with  $T$  the death time of a subject and, for any proposition  $B$ ,  $I(B)$  is the indicator function that takes the value 1 if  $B$  is true and zero otherwise. Thus  $I(T>t) = 1$  if a subject is alive at  $t$  and zero if dead at  $t$ . If  $I(T>t) = 0$ , I include in  $L(t)$  the exact day of death.

By convention, set  $A^*(t)$  and the remaining components of  $L(t)$  to zero once a subject has died.

The baseline covariates  $L(0)$  include covariate and BMI data on a subject before follow-up starts at the age of 18 years. Specifically, let  $BMI(0)$  denote BMI at (that is, just before) age 18 years (that is, time 0). Our inclusion of BMI just before age 18 years as a covariate rather than a treatment reflects the fact that ‘change’ in BMI since 18 years is our exposure. In particular, note that  $A^*(0)$  is the difference between BMI recorded just before 18 years and 1 month and BMI recorded just before 18 years. As is standard in the literature, I have taken change in BMI rather than change in weight in pounds as the exposure variable. Let  $\bar{A}^*(t)$  and  $\bar{L}^*(t)$  be the change in BMI and the covariate history through time  $t$  and  $\bar{A}^* = \bar{A}^*(K)$  be a subject’s (change in) BMI history through month  $K$  and  $\bar{L} = \bar{L}(K+1)$  be  $L$  history through the end of the study. A subject’s utility  $Y$ , a measure of quality-adjusted survival, is calculated from  $\bar{L} = \bar{L}(K+1)$  as  $\bar{L}$  includes the survival time of non-survivors, health status measures for survivors at the end of follow-up and time-varying health status factors.

### Potential for measured and unmeasured confounding

*Reverse causation and unmeasured confounding by subclinical disease.* In the literature on the effect of BMI on mortality, a controversy has arisen as whether and how to modify standard analytic methods to account for ‘reverse causation.’ Reverse causation refers to the well-accepted fact that preclinical (that is, undiagnosed) chronic disease, such as preclinical cancer, can cause both weight loss (or diminished weight gain) and death. It follows that among subjects with identical BMI history ( $\bar{A}^*(t-1)$ ,  $BMI(0)$ ) and measured covariate history  $\bar{L}(t)$  before age  $t$ , the subset whose monthly change  $A^*(t)$  in BMI is negative *are not comparable* with

regard to mortality risk to the subset with positive  $A^*(t)$ , even if BMI has no causal effect on mortality. That is, reverse causation implies unmeasured confounding by undiagnosed chronic disease. In fact, by an analogous argument, even among the subset with  $A^*(t)$  positive, there will be unmeasured confounding, because those with a small gain in BMI are more likely to have preclinical disease than those with a substantial gain.

It follows that one requires an analytic method that can adjust for unmeasured confounding due to the presence of preclinical disease. I will present a method that is appropriate under the additional assumption that we are able to specify an upper bound on the length of time a subject may have a subclinical illness severe enough to affect his weight, before that illness becomes clinically manifest.

*Measured confounders that are also intermediate variables.*

I next turn to the issue of confounding by measured factors, that is, by components of the covariate vector  $L(t)$ . For pedagogic purposes, in this subsection, it will be simpler to imagine that the unmeasured confounding due to reverse causation discussed above is not present. Now it is fairly well accepted that obesity causes increased BP, increased LDLs, Db and decreased exercise and these four factors may in turn cause increased mortality. Thus, these four variables are intermediate variables on the causal pathway from BMI to mortality. To prevent underestimation of the overall effect of BMI on mortality due to adjusting for intermediate variables, many analyses of the effect of BMI on mortality have failed to adjust for BP, LDL, Db or exercise in the analysis. However, such a decision can only be justified if these potential intermediate variables do not also confound the BMI–mortality relationship.

A sufficient condition for these intermediate variables to be also confounders is that, among subjects with identical BMI history ( $BMI(0)$ ,  $\bar{A}^*(t-1)$ ) until  $t$ , the subset whose change  $A^*(t)$  in BMI is non-positive is *not comparable* with regard to past BP, LDL, Db and exercise history to the subset with positive  $A^*(t)$ . Such non-comparability implies that, if data on time-varying BP, LDL, Db and exercise history are not used in the analysis, there will exist a non-causal association between an increase in BMI during month  $t$  and subsequent adverse *mortality*, even under the null hypothesis of no overall effect of BMI on mortality. Such non-comparability can occur whenever some or all of these intermediate variables are either a cause of a change in BMI or are correlated with an unmeasured cause of a change in BMI.

For example, it is likely that lack of exercise causes weight gain. In that case, if regular exercise causes decreased mortality, then, in an analysis that fails to adjust for exercise history prior to  $t$ , the association between an increase of  $A^*(t)$  in BMI during month  $t$  and subsequent adverse *mortality* will be an overestimate of the true causal effect of  $A^*(t)$  on mortality, due to uncontrolled confounding by exercise.

Similarly, suppose that chronic emotional stress and low-grade depression not only cause weight gain by inducing overeating as a soothing, self-medicating behavior but also directly cause elevated BP, elevated LDL and Db independently through various stress-induced metabolic, immune and sympathetic nervous system effects. If, as is true in most observational databases, data on chronic emotional stress and low-grade depression are not recorded (that is, measured), then, even under the null hypothesis of no overall effect of BMI on mortality, the association between an increase in BMI and subsequent adverse *mortality* will tend to be positive, whether or not one adjusts for elevated BP, elevated LDL and Db in the analysis, due to uncontrolled confounding by chronic emotional stress and low-grade depression. However, these variables should be appropriately adjusted for in the analysis, because the magnitude of positive overestimation will often be much less if they are adjusted for, because of their correlation with the unmeasured causal confounder—chronic emotional stress and depression.

In contrast with the last paragraph, suppose there is no confounding by chronic emotional stress and low-grade depression; rather, in the observational database, most individuals who developed an elevated BP, elevated LDL or Db became concerned about their health and instituted a diet that resulted in their gaining less weight than those without these conditions. Then the association found between an increase in BMI and subsequent adverse *mortality* in an analysis that fails to adjust for these variables at  $t$  would tend to underestimate the true causal effect of BMI on mortality due to negative confounding. I conclude that elevated BP, elevated LDL and Db could confound the association between increase in BMI and subsequent adverse *mortality* in either a negative or positive direction, depending on which of the mechanisms described in this paragraph and the last two predominate.

In summary, time-dependent covariates such as exercise (that is, physical activity), BP, LDL or Db that are recorded in  $L(t)$  may be both intermediate variables on the causal pathway from BMI to death and confounders of the BMI–death relationship. It follows that one requires an analytic method that can appropriately adjust for the effects of measured time-varying covariates that are simultaneously intermediate variables and time-dependent confounders.

*Intractable unmeasured confounding in subgroups.* There may be subgroups defined by measured variables in whom confounding by unmeasured factors is intractable. For example, among persons with diagnosed chronic renal, liver, pulmonary or cardiac disease, rapid weight gain can indicate increasing edema (water retention) due to unmeasured disease progression rather than increasing fat stores; as a consequence, among chronic disease patients with identical pasts, comparability would not hold because individuals experiencing rapid weight gain may be at increased risk of

death due to unmeasured progression of disease compared to those with lesser weight gain. In such a case, unmeasured confounding by disease progression may be intractable.

Using other arguments, various investigators have argued that in both the subgroup of subjects over age 70 years and the subgroup with BMI less than 21 kg/m<sup>2</sup>, subjects gaining weight at different rates are not comparable owing to unmeasured confounding factors, even when data have been collected on many potential confounders.

Therefore, one needs an analytic method that can remain valid even when there exists intractable confounding among subjects with a diagnosed chronic disease, an age of greater than 70 years, or a BMI below 21 kg/m<sup>2</sup>. In the next section, I describe an analytic method that satisfies the requirements of this and the two previous subsections.

#### *A simplified description of g-estimation of SNMs*

In this subsection, I give a non-technical, conceptual description of how, even in the presence of the measured and unmeasured confounding described in the section Potential for measured and unmeasured confounding, g-estimation of SNMs can be used to estimate the expected utility had, in contrast to the fact, all non-smoking 18-year-old American men in 1950 been put on a stringent mandatory diet that guaranteed that no one would ever weigh more than their weight at the age of 18 years. To avoid technical digressions and thereby keep the description centered on important conceptual issues, this non-technical description is neither complete nor fully accurate. Section Estimation of the effect of the 'maintain baseline weight intervention' onwards provides a complete and accurate description. This completeness and accuracy unfortunately place greater technical demands on the reader.

A locally RP SNM for  $Y$  is a rule that takes as input a subject's observed utility  $Y$ , their observed BMI and covariate history through the end of the study, and an unknown parameter  $\beta^*$  and outputs the utility  $Y_0$  that would have been observed if, possibly in contrast to the fact, the subject had followed the dietary intervention of the first paragraph of the Introduction. If the rule is correct and we knew the value of  $\beta^*$  then we could calculate  $Y_0$  for each study subject. The average of these  $Y_0$  in the cohort of all non-smoking 18-year-old American men in 1950 is our quantity of interest: the expected (that is, average) utility had one implemented a dietary intervention that guaranteed that no one would ever weigh more than they did at the age of 18 years. However, we do not know the value of  $\beta^*$ . Thus, the challenge is to estimate  $\beta^*$  from the data. When, as in the section Measured confounders that are also intermediate variables, all confounding is due to measured variables, Robins<sup>2</sup> proposed a method of estimation called g-estimation that is described next.

Let us define  $A(t)$  to be the difference between a subject's observed BMI,  $\text{BMI}(t+1)$ , just prior to month  $t+1$  and his maximum value  $\text{BMI}_{\max}(t)$  of BMI prior to month  $t$ , whenever that difference is non-negative. When the differ-

ence is negative, we simply set  $A(t)$  to be zero. Formally then

$$A(t) = \text{BMI}(t+1) - \text{BMI}_{\max}(t) \quad \text{if } \text{BMI}(t+1) \geq \text{BMI}_{\max}(t) \quad (2)$$

$$A(t) = 0 \quad \text{if } \text{BMI}(t+1) < \text{BMI}_{\max}(t). \quad (3)$$

$A(t)$  is non-negative.

If the only confounding is due to measured factors, then among subjects with the same BMI and covariate history prior to time  $t$  with positive  $A(t)$ , the increase of  $A(t)$  in BMI between  $t$  and  $t+1$  will be conditionally uncorrelated with  $Y_0$ . Thus to estimate  $\beta^*$ , we simply try many different guesses  $\beta$ . If a particular guess  $\beta$  were the true  $\beta^*$ , then the output of the rule would be uncorrelated with  $A(t)$ . Thus I choose as our estimate  $\hat{\beta}$  of  $\beta^*$ , the guess  $\beta$  that results in an output that has smallest conditional correlation with  $A(t)$  when we combine the information across all months  $t$  from 0 to the end of follow-up at  $K+1$ .

When as in the section Intractable confounding in subgroups, there are certain identifiable subgroups in whom confounding is intractable, bias can result because the output of the rule will be conditionally correlated with  $A(t)$  even when  $\beta = \beta^*$ . To eliminate this bias, it suffices to search for lack of correlation with  $A(t)$  only among subjects who are not in any of these intractable confounded subgroups at time  $t$ . That is, I simply restrict our g-estimation procedure at a given time  $t$  to subjects who are not currently members of any intractably confounded subgroup as in Joffe *et al.*<sup>16</sup>

When there is unmeasured confounding by subclinical disease such as in the section Reverse causation and unmeasured confounding by subclinical disease, I must modify our g-estimation procedure. Suppose one can specify an upper bound, say 6 years, on the length of time a subject may have a subclinical illness severe enough to affect weight gain, before that illness becomes clinically manifest. Then one can still validly estimate  $\beta^*$  if one restricts the g-estimation procedure at a given time  $t$  to those subjects with positive  $A(t)$  who would have remained alive and free of a diagnosed chronic (that is, of clinical) disease for the 6 years following  $t$  had, possibly in contrast to the fact, they followed a diet that prevented any further weight gain over those 6 years; by our assumption of a 6-year upper bound, such subjects did not have their weight gain affected by an undiagnosed chronic disease. It does not suffice to restrict to subjects who actually remained alive and free of clinical disease for the 6 years following  $t$ , because if BMI change  $A(t)$  at  $t$  causally effects the onset of clinical disease and/or survival in the following 6 years, the variable 'survival without clinical disease for 6 years after  $t$ ' is a response affected by the exposure  $A(t)$  and thus cannot be adjusted for without introducing selection bias as explained in Hernán *et al.*<sup>6</sup> Thus to validly estimate  $\beta^*$  using g-estimation, one must be able to determine those 'subjects who would have remained alive and free of clinical disease for the 6 years following  $t$  had, possibly contrary to fact, they followed a diet that prevented any further weight gain over those 6 years.'

One can do so by specifying a second SNM, called a locally RP SNFTM, for the effect of change in BMI on the time  $X$  to the diagnosis of chronic disease or death (whichever comes first). A locally rank-preserving SNFTM is a rule that takes as input a subject's observed time  $X$  to (the earlier of) death or a diagnosed chronic disease, their observed BMI and covariate history through the end of the study, and an unknown parameter  $\psi^*$ , and a time  $t$  and outputs the time  $X_t$  that would have been observed if, possibly in contrast to the fact, the subject had followed a dietary intervention in which no further weight was gained after time  $t$ . If  $\psi^*$  were known or well estimated, we could compute  $X_t$  for each subject, determine which subjects'  $X_t$  failed to exceed  $t$  by more than 6 years, and exclude such subjects from the g-estimation procedure used to estimate  $\beta^*$ .

Thus, it only remains to estimate the parameter  $\psi^*$  of our locally rank-preserving SNFTM in the presence of unmeasured confounding by subclinical disease. Now among subjects' with the same BMI and covariate history prior to time  $t$  with positive  $A(t)$  who are not members of an identifiable subgroup with intractable confounding, the change  $A(t)$  in BMI between  $t$  and  $t+1$  will be uncorrelated with  $X_t$  if we restrict to subjects with  $X_t$  exceeding  $t$  by more than 6 years. Thus to estimate  $\psi^*$ , I simply try many different guesses  $\psi$ . If a particular guess  $\psi$  were the true  $\psi^*$ , the output of the SNFTM rule would be uncorrelated with  $A(t)$  when I restrict the g-estimation procedure to subjects whose output exceeds  $t$  by more than 6 years. Thus, I choose as the estimate  $\hat{\psi}$  of  $\psi^*$ , the guess  $\psi$  that results in an output that, under this restricted g-estimation procedure, has the smallest conditional correlation with  $A(t)$  when I combine the information across all times  $t$ .

Before proceeding to the more technical part of the paper, I provide a brief non-technical discussion of several important but subtle points about SNMs. First, locally rank-preserving SNMs assume that the effect of a given increase in BMI on the utility  $Y$  and on  $X$  is the same for any two subjects with the same past measured covariate history. This assumption is biologically implausible as unmeasured genetic and environmental factors will clearly modify the magnitude of the effect of weight gain on the responses  $Y$  and  $X$ . Fortunately, we prove in the section Estimation of the effect of the 'maintain baseline weight intervention' that our g-estimator of the mean of  $Y_0$  remains valid even if we allow the magnitude of the effect of weight gain on  $Y$  and  $X$  to be modified in an arbitrary manner by unmeasured genetic and environmental factors.

The description of g-estimation of the parameters  $\psi^*$  of our SNFTM model for  $X$  assumed that the time  $X$  to death or diagnosed chronic disease was available for every study subject. However, by the end of follow-up, a number of study subjects will remain alive and free of chronic disease. Such subjects are said to be censored. In the section Censoring, I show how our g-estimation procedures can be modified to appropriately account for these censored observations.

The estimate of the mean of  $Y_0$  will be biased if either the SNMM for  $Y$  or the SNFTM for  $X$  are misspecified. I discuss below how to construct tests for misspecification. However, I also show that the power of such tests to detect model misspecification can be quite limited in the presence of reverse causation by subclinical disease and intractable confounding in identifiable subgroups. In Section Intractable confounding in subgroups, I offer some suggestions on how the impact of this limited power on the quality of one's inferences can be lessened if one is willing to change the parameter that is being estimated.

### Estimation of the effect of the 'maintain baseline weight intervention'

In this section, I describe how we can use g-estimation of SNMs to estimate the expected utility had one put all non-smoking 18-year-old American men on a stringent mandatory diet that guaranteed that no one would ever weigh more than their baseline weight established at the age of 18 years. For pedagogic reasons, I first consider the simpler setting in which there is no unmeasured confounding by preclinical disease.

#### Case 1: no unmeasured confounding by preclinical disease

*A locally rank-preserving SNM.* An SNM is a model for counterfactual variables  $Y_m$  that denote a subject's utility measured at end of follow-up under the following counterfactual dietary intervention:

*Time  $m$  dietary intervention:* The subject follows his observed diet up to month  $m$  following his eighteenth birthday and, from month  $m$  onwards, the subject is weighed every day: (i) whenever his weight is greater than or equal to his maximum monthly BMI up to  $m$  (that is,  $\text{BMI}_{\max}(m) \equiv \max\{\text{BMI}(0), \dots, \text{BMI}(m)\}$ ), the subject's caloric intake is restricted until the subject's BMI falls to below  $\text{BMI}_{\max}(m)$ ; (ii) whenever his weight is less than  $\text{BMI}_{\max}(m)$ , the subject is allowed to eat as he pleases without any intervention.

A subject's responses had, possibly in contrast to the fact, he been made to follow a time  $m$  dietary intervention are referred to as counterfactual responses. We assume that  $Y_m$  is well-defined in the sense that its value is insensitive to the unspecified details of exactly how the subject's calories are to be restricted in (i). We also assume these counterfactual responses are observed only for those  $m$  for which a subject's actual BMI history was consistent with his having followed the time  $m$  dietary intervention. For other values of  $m$ , the time  $m$ -specific counterfactuals remain unobserved.

The time 0 dietary intervention is the dietary intervention in the first paragraph of the Introduction. The counterfactual  $Y_0$  is the utility corresponding to this regimen. Thus, the expected value  $E[Y_0]$  of  $Y_0$  is our parameter of interest: the expected utility had we placed in 1950 all non-smoking

18-year-old American men on a diet that guaranteed that no one would ever weigh more than they did at the age of 18 years.

Note that  $Y_{K+1} = Y$ : if one were to follow his actual observed diet up to the time  $K+1$  at which the study ends, then no dietary intervention would have occurred. Hence, the counterfactual  $Y_{K+1}$  must be the observed (that is, actual)  $Y$ .

By definition, a subject's observed data through  $k$  (but before  $k+1$ ) is inconsistent or incompatible with following the 'time  $m$  dietary intervention' if and only if  $\text{BMI}(k+1) > \text{BMI}_{\max}(m)$  for some  $k \geq m$ .

It follows that it is only when the individual's observed data through time  $m$  are incompatible with the 'time  $m$  dietary intervention' is  $A(m) \neq 0$ . If an individual's observed data are consistent with his having followed the 'time  $m$  dietary intervention,' it is consistent with his having followed the 'time  $t$  dietary intervention' for  $t > m$ .

Note that  $Y_{m+1} - Y_m = 0$  whenever  $A(m) = 0$ . If  $A(m) \neq 0$ ,  $Y_{m+1} - Y_m$  is the difference between (i) a subject's utility when he has his observed  $\overline{\text{BMI}}(m+1)$  history and thereafter, possibly in contrast to the fact, the subject follows the dietary intervention that guarantees his  $\text{BMI}(k)$  for  $k > m+1$  never again exceeds  $\text{BMI}_{\max}(m+1)$  and (ii) his utility when he has his observed  $\overline{\text{BMI}}(m)$  history and thereafter, possibly in contrast to the fact, the subject follows the dietary intervention that guarantees his BMI at  $m+1$  equals his observed  $\text{BMI}_{\max}(m)$  (rather than his observed BMI at  $m+1$ ) and that his  $\text{BMI}(k)$  for  $k > m+1$  never again exceeds  $\text{BMI}_{\max}(m)$ . As a kind of shorthand for the previous sentence, whenever  $A(m) \neq 0$ , we will refer to  $Y_{m+1} - Y_m$  as the causal effect of a final blip of exposure of magnitude  $A(m)$  on the subject's utility.

An additive locally rank-preserving SNM is a deterministic model for the magnitude of the effect of a treatment  $A(m)$  on  $Y_{m+1} - Y_m$ . Mathematically, an additive locally rank-preserving SNM assumes that for each time  $m = 0, \dots, K$ ,

$$Y_{m+1} - Y_m = \gamma_m[A(m), \overline{A}(m-1), \overline{L}(m), \beta^*] \quad (4)$$

where (i)  $\beta^*$  is the unknown true parameter vector, and (ii)  $\gamma_m[A(m), \overline{A}(m-1), \overline{L}(m), \beta]$  is a known function [such as  $\{\beta_0 + \beta_1 m + \beta_2^T L(m)\} A(m)$ ] satisfying the restrictions  $\gamma_m[A(m), \overline{A}(m-1), \overline{L}(m), \beta] = 0$  if  $A(m) = 0$  or  $\beta = 0$ . Here,  $\beta_2$  is a column vector of length equal to that of the vector  $L(m)$ . Furthermore,  $T$  as a superscript denotes the transpose of a matrix or vector. The first restriction must logically hold because, by definition, if  $A(m) = 0$ ,  $Y_{m+1} = Y_m$ . We now show that the second restriction guarantees that  $\beta^* = 0$  encodes the sharp null hypothesis that 'following a diet that prevents one's BMI from ever exceeding the baseline BMI' has no effect on any subject's utility.

Recalling that  $Y_{K+1} = Y$ , the model (4) is seen to be equivalent to the model

$$Y_m = Y - \sum_{j=m}^K \gamma_m[A(j), \overline{A}(j-1), \overline{L}(j), \beta^*] \quad (5)$$

for  $m = 0, 1, \dots, K$ . To help understand Equation (5), consider first the special case  $m = K$ . Then Equation (5) says that to calculate  $Y_K$  from  $Y$ , we remove the causal effect  $\gamma_K[A(K), \overline{A}(K-1), \overline{L}(K), \beta^*]$  of exposure  $A(K)$  at the last time  $K$ . Next consider the special case  $m = 0$ . Then Equation (5) says that to calculate  $Y_0$ , one successively removes the effect of exposure at times  $K, K-1, \dots, 0$ . It follows from the restriction  $\gamma_m[A(m), \overline{A}(m-1), \overline{L}(m), 0] = 0$ , that  $\beta^* = 0$  implies  $\gamma_m[A(m), \overline{A}(m-1), \overline{L}(m), \beta^*] = 0$  for each  $m$ . Thus  $\beta^* = 0$  encodes the sharp null hypothesis that  $Y_0 = Y_m = Y$  for all subjects and all  $m$ . In other words, one's utility at the end of the study will be the same regardless of whether or not one follows any 'time  $m$  dietary intervention.'

As, by Equation (5), a locally rank-preserving SNM directly maps an individual's observed utility  $Y$  to the utility an individual would have under the 'time  $m$  dietary intervention,' it is a model for individual causal effects.

Possible choices of  $\gamma_m[a(m), \overline{a}(m-1), \overline{l}(m), \beta]$  include (i)  $\beta a(m)$ , (ii)  $(\beta_0 + \beta_1 m) a(m)$ , (iii)  $\{\beta_0 + \beta_1 m + \beta_2^T l(m)\} a(m)$ . In model (i), the effect of a change of  $A(m)$  in BMI is the same for all  $m$ . Under model (ii), the effect varies linearly with time  $m$ . Under model (iii), the causal effect of  $A(m)$  is modified by the most recent covariate history.

In the following, we assume the observed data  $O$  on each subject is  $O = (Y, \overline{L}, \overline{A}) \equiv (Y, \overline{L}(K+1), \overline{A}(K))$ . That is  $O$  consists of a subject's utility  $Y$  and his covariate and treatment histories through the end of the study. The inclusion of  $\overline{A}(K)$  is actually redundant, as the  $A$ -history  $\overline{A}(K)$  is determined by  $\overline{\text{BMI}}(K+1)$ , and  $\overline{\text{BMI}}(K+1)$  is a component of  $\overline{L}(K+1)$ . Thus, we could write the observed data as simply  $(Y, \overline{L}(K+1))$ . However, because we wish to use results on g-estimation of SNMs that were derived in previous papers in which  $\overline{A}(K)$  was not determined by  $\overline{L}(K+1)$ , we will continue to write  $O = (Y, \overline{L}, \overline{A})$  and accept some redundancy in the notation. Let

$$Y_m(\beta) = Y - \sum_{j=m}^K \gamma_m[A(j), \overline{A}(j-1), \overline{L}(j), \beta] \quad (6)$$

so, under our model,  $Y_m = Y_m(\beta^*)$ . Note that, for each  $\beta$ ,  $Y_m(\beta)$  can be computed from the observed data  $(Y, \overline{L}, \overline{A})$ . Suppose we had a consistent estimate  $\hat{\beta}$  of  $\beta^*$ . Then  $Y_0(\hat{\beta})$  would be a consistent estimate of  $Y_0 = Y_0(\beta^*)$ . Thus, the average  $\sum_{i=1}^n Y_{0i}(\hat{\beta})/n$  over the  $n$  study subjects would be a consistent estimate of the parameter of interest  $E[Y_0]$ . Further  $\sum_{i=1}^n Y_{0i}(\hat{\beta})/n - \sum_{i=1}^n Y_i/n$  would be a consistent estimate of the difference  $E[Y_0] - E[Y]$  between the expected utility  $E[Y_0]$  under a dietary intervention guaranteeing BMI never exceeds the baseline BMI and the expected utility  $E[Y]$  in the absence of any dietary intervention. We show below how one can obtain a consistent estimate  $\hat{\beta}$  by g-estimation if a certain comparability assumption holds.

*The innovative aspect of our SNM:* The most important and innovative aspect of our model is that it models the causal effect on the utility of an increase in BMI of  $A(m)$  over a subject's maximum past BMI,  $\text{BMI}_{\max}(m)$ . It does not model

and thus is agnostic about the causal effect (a) of any decrease in BMI or (b) of any increase in BMI between  $m$  and  $m+1$  that fails to result in one's BMI exceeding  $\text{BMI}_{\max}(m)$ . Our model (4) is thus more robust than alternative models that would also model (a) or (b), because our model makes fewer assumptions than such alternative models, and thus is less likely to be misspecified. However, the small number of assumptions made by our model is sufficient for our purposes; if we can consistently estimate the parameter  $\beta^*$ , we can consistently estimate our parameter of interest  $E[Y_0]$ .

Thus, it only remains to estimate  $\beta^*$ . In this section, we will do so under the following assumption, which will be weakened in later sections. Define the indicator variable  $\Xi(m)$  taking values in the two-element set  $\{0, 1\}$  by

$$\Xi(m) = 1 \Leftrightarrow \text{BMI}(m+1) \geq \text{BMI}_{\max}(m) \quad (7)$$

That is  $\Xi(m)$  takes the value 1 if a subject's BMI just before  $m+1$  is at least as great as his maximum BMI up to time  $m$ . Otherwise  $\Xi(m)$  takes the value 0.

*Comparability assumption (CO):* Among subjects with the same  $\bar{A}(m-1)$  history and covariate history  $\bar{L}(m)$  (which includes BMI history  $\bar{\text{BMI}}(m)$ ) and with  $\Xi(m)=1$ ,  $A(m)$  is statistically independent of the counterfactual  $Y_m$ . Formally, conditional on  $(\bar{A}(m-1), \bar{L}(m), \Xi(m)=1)$ ,  $A(m)$  is independent of  $Y_m$ . (As past A-history  $\bar{A}(m-1)$  is determined by  $\bar{\text{BMI}}(m), \bar{A}(m-1)$  in the conditioning event  $(\bar{A}(m-1), \bar{L}(m))$  is redundant; nonetheless we shall retain the  $\bar{A}(m-1)$ .)

A comparability assumption such as CO is often referred to as an assumption of no confounding by unmeasured factors or as an assumption of sequential randomization.

*Remark:* To understand why we conditioned on  $\Xi(m)=1$  in the CO assumption, imagine we had instead assumed that  $A(m)$  is independent of  $Y_m$  conditional on  $(\bar{A}(m-1), \bar{L}(m))$ . This would have implied that among the subset of subjects with a given  $(\bar{A}(m-1), \bar{L}(m))$ , the subgroup with  $A(m) \neq 0$  would have the same distribution of the utility  $Y_m$  under the time  $m$ -dietary intervention as the subgroup with  $A(m)=0$ . But, under the time  $m$ -intervention, all subjects in the  $A(m) \neq 0$  subgroup would have  $\text{BMI}(m+1)$  equal to their common  $\text{BMI}_{\max}(m) \in \bar{L}(m)$ , whereas many subjects with  $A(m)=0$  (specifically, those with  $\Xi(m)=0$ ) would have  $\text{BMI}(m+1) < \text{BMI}_{\max}(m)$ . Thus, the  $A(m)=0$  subgroup will have lower BMI at  $m+1$  than the  $A(m) \neq 0$  subgroup under the time  $m$ -intervention. Suppose the null hypothesis of no biological BMI effect is false. Then, for an individual with  $A(m)=0$ , their utility  $Y_m$  should depend on their BMI at  $m+1$ . As such, it extremely unlikely that the  $A(m)=0$  and  $A(m) \neq 0$  subgroups would be comparable. In contrast, if, as in assumption CO, we restrict the  $A(m) \neq 0$  subgroup to a subset of the subjects with  $\Xi(m)=1$ , then given  $\bar{L}(m)$ , this restricted  $A(m)=0$  subgroup, like the  $A(m) \neq 0$  subgroup, will have  $\text{BMI}(m+1)$  equal to the common  $\text{BMI}_{\max}(m)$  under the intervention, so the assumption of non-comparability is plausible.

Under the CO assumption, we can obtain a consistent estimator of  $\beta^*$  by g-estimation as follows. We specify a linear regression model.

$E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m)=1] = \alpha^T W(m)$  for  $m=0, \dots, K$ . Here  $W(m) = w_m[\bar{L}(m), \bar{A}(m-1)]$  is a vector of covariates calculated from a subject's past data,  $\alpha^T$  is a row vector of unknown parameters, and each person-month is treated as an independent observation, so each person contributes up to  $K+1$  observations. However, person months for which  $\Xi(m) \neq 1$  are excluded from the regression. Examples of  $W(m) = w_m[\bar{L}(m), \bar{A}(m-1)]$  would be the transpose of the row vector  $(m, L^T(m), L^T(m-1))$ . Let  $\hat{\alpha}$  be the ordinary least squares (OLS) estimator of  $\alpha$  computed using a standard statistical package.

For the moment assume  $\beta$  is one-dimensional (1D). Let  $\beta_{\text{low}}$  and  $\beta_{\text{up}}$  be much smaller and larger, respectively, than any substantively plausible value of  $\beta^*$ .

Then, separately, for each  $\beta$  on a grid from  $\beta_{\text{low}}$  to  $\beta_{\text{up}}$ , say  $\beta_{\text{low}}, \beta_{\text{low}}+0.1, \beta_{\text{low}}+0.2, \dots, \beta_{\text{up}}$ , perform the score test of the hypothesis  $\theta=0$  in the extended linear model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), Y_m(\beta), \Xi(m)=1] = \alpha^T W(m) + \theta Y_m(\beta) \quad (8)$$

that adds the covariate  $Y_m(\beta)$  at each time  $m$  to the above (pooled over persons and time) linear model. A 95% confidence interval for  $\beta^*$  is the set of  $\beta$  for which an  $\alpha=0.05$  two-sided score test of the hypothesis  $\theta=0$  does not reject. The g-estimate  $\hat{\beta}$  of  $\beta^*$  is the value of  $\beta$  for which the score test takes the value zero (that is, the  $P$ -value is one).

The validity of g-estimation is proved as follows. By our comparability assumption,  $Y_m(\beta^*)$  and  $A(m)$  are conditionally independent given  $(\bar{L}(m), \bar{A}(m-1), \Xi(m)=1)$ . That is,  $Y_m(\beta^*)$  is not a predictor of  $A(m)$  given  $(\bar{L}(m), \bar{A}(m-1), \Xi(m)=1)$ , which implies that the coefficient  $\theta$  of  $Y_m(\beta)$  must be zero in the extended model when  $\beta=\beta^*$ , provided the model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m)=1] = \alpha^T W(m)$$

is correctly specified.

Now, we do not know the true value of  $\beta$ . Therefore, any value  $\beta$  for which the data are consistent with the parameter  $\theta$  of the term  $\theta Y_m(\beta)$  being zero might be the true  $\beta^*$ , and thus belongs in our confidence interval. If consistency with the data is defined at the 0.05 level, then our confidence interval will have coverage of 95%. Furthermore, the g-estimate  $\hat{\beta}$  of  $\beta^*$  is that  $\beta$  for which adding the term  $\theta Y_m(\beta)$  does not help to predict  $A(m)$  whatsoever, which is the  $\beta$  for which the score test of  $\theta=0$  is precisely zero. The g-estimate  $\hat{\beta}$  is also the value of  $\beta$  for which the OLS estimator of  $\theta$  is precisely zero.

It may appear peculiar that a function  $Y_m(\beta)$  of the response  $Y$  measured at the end of follow-up is being used to predict  $A(m)$  at earlier times. However, this peculiarity evaporates when one recalls that, for each  $\beta$  on our grid, we are testing the null hypothesis that  $\beta=\beta^*$ , and, under this null,  $Y_m(\beta)$  is the counterfactual  $Y_m$ , which we can view as already existing at time  $m$  (although we cannot observe its

value until time  $K + 1$  and then only if  $A(t)$  in the observed data is zero from  $m$  onwards).

Suppose next that the parameter  $\beta$  is a vector. To be concrete, suppose we consider the model with

$$\gamma_m[a(m), \bar{a}(m-1), \bar{l}(m), \beta] = a(m)\{\beta_0 + \beta_1 m + \beta_3^T l(m)\}$$

so  $\beta$  is of dimension  $\dim(l(m)) + 2$  where  $\dim(l(m))$  is the dimension of  $l(m)$  which, for concreteness, we take to be 3. Hence  $\beta$  is 5D. Then we would use a 5D grid, one dimension for each component of  $\beta$ . So if we had 20 grid points for each component, we would have  $20^5$  different values of  $\beta$  on our 5D grid. Now to estimate five parameters one requires five additional covariates. Specifically, let  $Q_m(\beta) = q_m[\bar{L}(m), \bar{A}(m-1), Y_m(\beta)]$  be a 5D vector of functions of  $(\bar{L}(m), \bar{A}(m-1), Y_m(\beta))$ , such as  $Q_m(\beta) = [1, m, L^T(m)](Y_m(\beta))$ . We use the extended model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), Y_m(\beta), \Xi(m) = 1] = \alpha^T W(m) + \theta^T Q_m(\beta)$$

Our g-estimate  $\hat{\beta}$  is the  $\beta$  for which the 5 degree of freedom score test that all five components of  $\theta$  equal zero is precisely zero. The particular choice of the functions  $q_m$  does not affect the consistency of the point estimate, but it affects the width of the confidence interval.

When  $\gamma_m[a(m), \bar{A}(m-1), \bar{L}(m), \beta] = a(m)\beta^T R_m$  is linear in  $\beta$  with  $R_m = r_m(\bar{L}(m), \bar{A}(m-1))$  being a vector of known functions and we choose  $Q_m(\beta) = Q_m^* Y_m(\beta)$  linear in  $Y_m(\beta)$ , then, given the OLS estimator  $\hat{\alpha}^T$  of  $\alpha^T$  in the model  $E[A(m)|\bar{L}(m), \bar{A}(m-1)] = \alpha^T W(m)$ , there is an explicit closed form expression for  $\hat{\beta}$  given by

$$\hat{\beta} = \left\{ \sum_{i=1, m=0}^{i=n, m=K} \Xi_i(m) G_{im}(\hat{\alpha}) Q_{im}^* S_{im}^T \right\}^{-1} \times \left\{ \sum_{i=1, m=0}^{i=n, m=K} \Xi_i(m) Y_i G_{im}(\hat{\alpha}) Q_{im}^* \right\} \quad (9)$$

with  $G_{im}(\hat{\alpha}) = [A_i(m) - \hat{\alpha}^T W_i(m)]$ ,  $S_{im} = \sum_{j=m}^K A_i(j) R_{ij}$ .

**Identification:** Suppose that two different values of  $\beta$ , say  $\hat{\beta}$  and  $\tilde{\beta}$ , both make the 5 degree of freedom score test precisely zero and yet the two confidence interval for  $\beta^*$  centered at  $\hat{\beta}$  and  $\tilde{\beta}$  do not overlap. How should we choose between the estimates? (In such a case, the matrix whose inverse is required in (9) will not be invertible and so (9) will fail.) As we can use any five vectors  $Q_m(\beta) = q_m[\bar{L}(m), \bar{A}(m-1), Y_m(\beta)]$  in our procedure, one simple approach is to try other choices of  $Q_m(\beta)$  until we find a  $Q_m(\beta)$  for which our confidence interval for  $\beta^*$  includes only one of the  $\hat{\beta}$  and  $\tilde{\beta}$ , declare the one included to be our point estimate of  $\beta^*$  and ignore the excluded one. Will this approach always succeed? In general, this approach should succeed in rather quickly excluding all but one of the values of  $\beta$  that originally made the score test zero, provided that the model  $\gamma_m[a(m), \bar{a}(m-1), \bar{l}(m), \beta]$  is correct, except when  $\beta^*$  is not identified. By definition  $\beta^*$  is not identified if there is a  $\beta^{**}$  different from the true parameter  $\beta^*$  such that, with an infinite sample size,  $\beta^{**}$ , like  $\beta^*$ , makes the 5 degree of freedom score test precisely zero for

all choices of  $Q_m(\beta)$ . In our model, it follows from Robins<sup>7,8</sup> that under the positivity assumption that

$$\Pr[A(m) = 0 | \bar{A}(m-1), \bar{L}(m), \Xi(m) = 1] \neq 0 \quad (10)$$

for all subjects and  $m = 0, \dots, K$ ,  $\beta^*$  is identified. In our context, the positivity assumption is a very weak assumption, which is almost certainly true. Hence, for the remainder of the paper, we will silently assume that it holds.

*Remark:* We considered a linear regression model for  $E[A(m)|\bar{L}(m), \bar{A}(m-1), Y_m(\beta), \Xi(m) = 1]$  in the above for expositional simplicity. In practice, as  $A(m) \geq 0$ , we might use a log linear model that specifies

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), Y_m(\beta), \Xi(m) = 1] = \exp\{\alpha^T W(m) + \theta^T Q_m(\beta)\}$$

and fit by nonlinear least squares. In that case, in Equation (9),  $G_{im}(\hat{\alpha}) = [A_i(m) - \exp\{\hat{\alpha}^T W_i(m)\}]$ . Alternatively, we could replace the response variable  $A(m)$  in the linear regression by  $\ln(A_m + 0.1)$  where the 0.1 is added to insure the logarithm remains finite even when  $A_m = 0$ . In that case,  $G_{im}(\hat{\alpha}) = [\ln\{A_i(m) + 0.1\} - \hat{\alpha}^T W_i(m)]$ .

*An additive SNMM.* An additive locally rank-preserving SNM<sup>8</sup> implies that if two subjects have the same observed data  $O = (Y, \bar{L}, \bar{A})$  they will have the same value of  $Y_0$  under the ‘time 0 dietary intervention’ of the Introduction. That is the model implies that for these subjects, the effect of the ‘time 0 dietary intervention’ will be identical. This assumption is clearly biologically implausible in view of between-subject heterogeneity in unmeasured genetic and environmental factors. To overcome this limitation, we consider an additive SNMM

$$E[Y_{m+1} - Y_m | \bar{A}(m), \bar{L}(m)] = \gamma_m[A(m), \bar{A}(m-1), \bar{L}(m), \beta^*] \quad (11)$$

that models the conditional mean of  $Y_{m+1} - Y_m$  given  $(\bar{A}(m), \bar{L}(m))$  rather than the individual differences  $Y_{m+1} - Y_m$ , and thus does not impose local rank preservation. In particular,  $Y_m$  no longer is equal to  $Y_m(\beta^*)$ . However, Robins<sup>8,9</sup> proved the additive SNMM implies (and, in fact is equivalent to) the assumption that  $Y_m$  and  $Y_m(\beta^*)$  have the same mean given,  $\bar{A}(m), \bar{L}(m), \Xi(m) = 1$ . That is

$$E[Y_m | \bar{A}(m), \bar{L}(m)] = E[Y_m(\beta^*) | \bar{A}(m), \bar{L}(m)] \quad (12)$$

for each  $m$ . Further he proved that, under the CO assumption, g-estimation of  $\beta^*$  retains all the properties described above, even in the absence of local rank preservation, except now the function  $Q_m(\beta)$  must be chosen linear in  $Y(\beta)$ , that is,  $Q_m(\beta) = Q_m^* Y_m(\beta)$  as above.

As a consequence, the definition of non-identifiability must be modified as follows: the parameter  $\beta^*$  of an SNMM  $\beta^*$  is not identified if there is a  $\beta^{**}$  different from the true parameter  $\beta^*$  such that, with an infinite sample size,  $\beta^{**}$ , like  $\beta^*$ , makes the 5 degree of freedom score test precisely zero for all choices of  $Q_m(\beta)$  that are linear in  $Y_m(\beta)$ . A detailed discussion of rank-preserving versus non-RP models occurs in the section Are remarkable results due to some sleight of hand.

*Alternative approaches:* Under the CO assumption, it is shown in Robins<sup>15</sup> that the  $E[Y_m]$  are non-parametrically identified for  $m=0, \dots, K$  from data  $O = (Y, \bar{L}, \bar{A})$  by the IPTW formula

$$E[YI\{\underline{A}(m) = \underline{0}(m)\}W(m)] \tag{13}$$

where  $\underline{A}(m) = (A(m), \dots, A(K))$  and the IPTW weight

$$W(m) = 1 / \prod_{k=m}^K \{\text{pr}(A(k) = 0 | \bar{A}(k-1), \bar{L}(k))\}^{\Xi(k)}$$

is the inverse of the conditional probability that a subject had his observed treatment  $\underline{A}(m) = \underline{0}(m)$ . That is,  $E[Y_m]$  is the weighted mean of the observed utility  $Y$  among subjects whose observed data were consistent with following the time  $m$  dietary intervention with weights given by the inverse of the conditional probability of having data consistent with following the intervention. Thus one could, in principle, consider estimating  $E[Y_0]$  non-parametrically by the weighted average of  $Y$  among subjects whose weight never exceeded their baseline weight at the age of 18 years with weights proportional to an estimate  $\widehat{W}(0)$  of  $W(0)$ . That is, by

$$\left[ \sum_{\{i: \underline{A}_i(0) = \underline{0}(0)\}} \widehat{W}_i(0) Y_i \right] / \left[ \sum_{\{i: \underline{A}_i(0) = \underline{0}(0)\}} \widehat{W}_i(0) \right]$$

The problem with this approach is that only the utility  $Y$  of the rare person whose weight never exceeds his age 18 weight contributes to the analysis. In contrast by specifying an SNMM, data on the utility  $Y$  of every subject contributes to the estimate of  $E[Y_0]$ . The price paid for the greater efficiency of an SNMM is the possibility of bias if the SNMM is misspecified.

However, under the CO,  $E[Y_m | \bar{A}(m), \bar{L}(m)] = E[Y_m | \bar{A}(m-1), \bar{L}(m)]$  and is non-parametrically identified by the formula

$$E[YI\{\underline{A}(m) = \underline{0}(m)\} / W(m) | \bar{L}(m), \bar{A}(m-1)]$$

Thus  $E[Y_{m+1} - Y_m | \bar{A}(m), \bar{L}(m)]$  is non-parametrically identified for  $m=0, \dots, K$ . Hence, given a sufficiently large sample size, one could in principle construct misspecification tests of the model (11) that have power against all alternatives when the model is incorrect. In practice, the available sample size may greatly limit the power to detect model misspecification.

IPTW estimation of marginal structural models and the parametric g-formula are alternative approaches to model-based estimation of  $E[Y_0]$  that also use data on every subject's utility  $Y$ . See Appendix 2 for further discussion.

*Remark:* The reader familiar with IPTW expects  $W(m)$  to be defined as

$$W(m) = 1 / \prod_{k=m}^K \{\text{pr}(A(k) = 0 | \bar{A}(k-1) = \underline{0}(k-1), \bar{L}(k))\}$$

rather than as

$$1 / \prod_{k=m}^K \{\text{pr}(A(k) = 0 | \bar{A}(k-1) = \bar{0}(k-1), \bar{L}(k))\}^{\Xi(k)}$$

In fact, the two expressions are equal if, without loss of generality, we adopt the coding convention that the vector

$L(k)$  includes  $\Xi(k)$  as a component because, under this coding,  $\text{pr}(A(k) = 0 | \bar{A}(k-1) = \bar{0}(k-1), \bar{L}(k))$  is one whenever  $\Xi(k)$  takes the value zero.

*Case 2: unmeasured confounding by preclinical disease*

In this section, we no longer assume  $A(m)$  is statistically independent of  $Y_m$  given  $\bar{A}(m-1), \bar{L}(m), \Xi(m) = 1$ . To describe our new comparability assumption, we need to introduce some further notation. Let  $X = \min(T, D)$  be the minimum of the time  $T$  to death and the time  $D$  to the diagnosis of a chronic disease, such as cancer, severe emphysema, liver or renal disease, or any other chronic condition that would be severe enough to affect weight gain. At each time  $m$ , the indicator  $I(X \leq m)$  is a component of  $L(m)$ . Further if  $I(X \leq m) = 1$ , the exact time  $X$  is observed and included in  $L(m)$ . Thus  $X$  is observed if  $X$  is less than  $K+1$ . However  $X$ , is censored (that is, not observed) on subjects whose  $X$  exceeds  $K+1$ , the end of follow-up time. For the present, we shall avoid the additional complications that arise from censoring by assuming that  $X$  is less than  $K+1$  for all subjects, so that the data  $O = (Y, X, \bar{L}, \bar{A})$  are observed on each subject. In the section Censoring, we relax this assumption and allow for censoring.

Let  $X_m = \min(T_m, D_m)$  be the counterfactual version of  $X = \min(T, D)$  had 'the time  $m$  dietary intervention' been carried out. Then we make the following more realistic assumption.

*Realistic comparability (RC) assumption:*  $A(m)$  is statistically independent of  $(Y_m, X_m)$  given  $\Xi(m) = 1, \bar{L}(m), \bar{A}(m-1)$  and  $\bar{U}(m) = \bar{0}(m)$ , where  $U(m) = 1$  if a subject has at  $m$  or had prior to  $m$ , an undiagnosed chronic disease that was sufficiently advanced to interfere with his normal weight trajectory. Otherwise  $U(m) = 0$ . We also assume  $U(m) = 1$  for subjects alive at  $m$  with  $X < m$  under the theory that there probably was a subclinical period prior to the time  $X$  of clinical diagnosis in which weight gain may have been altered. Note that  $U(m) = 0$  implies  $\bar{U}(m) = (U(0), \dots, U(m)) = \bar{0}(m)$  is also zero.

*Remark:* The RC assumption cannot be recast as  $(Y_m, X_m)$  independent of  $A(m)$  given  $(\bar{L}(m), \bar{A}(m-1), \bar{U}(m))$  even had we used the coding convention that vector  $L(k)$  includes  $\Xi(k)$  as a component, because, even under this coding,  $\text{pr}(A(m) = 0 | \bar{A}(m-1) = \bar{0}(m-1), \bar{L}(m), \bar{U}(m), (Y_m, X_m))$  would be neither zero nor one and could, under the RC assumption, depend on  $(Y_m, X_m)$  whenever  $\bar{U}(m) \neq 0$  and  $\Xi(m) = 1$ . For this reason it would perhaps be more precise to refer to the RC as a selective comparability assumption as it only implies comparability for a selected subset of the population.

We observe  $(Y, X, \bar{L}, \bar{A})$  but  $\bar{U} = (U(0), \dots, U(K))$  is, of course, generally unobserved. Thus,  $\bar{U}$  is an unmeasured confounder. The most crucial of several assumptions needed to allow consistent estimation of the parameter of interest  $E[Y_0]$  in this setting is the following.

*Clinical detection (CD) assumption:* Any subject who has  $U(m) = 1$  (that is, a sufficiently advanced undiagnosed

chronic disease at (or before)  $m$ ) and thereafter follows ‘the time  $m$  dietary intervention’ will either have died or been diagnosed with clinical chronic disease by time  $m + \zeta$ , where  $\zeta$  is assumed known. Formally

$$U(m) = 1 \Rightarrow X_m \leq m + \zeta \quad (14)$$

or equivalently

$$X_m > m + \zeta \Rightarrow \bar{U}(m) = \bar{0}(m) \quad (15)$$

Here  $\Rightarrow$  is translated as ‘implies.’ A typical choice for  $\zeta$  might be 72 months. It is useful to choose  $\zeta$  to be the minimal time for which (15) holds as this increases both the efficiency of g-estimators and the power of goodness-of-fit tests to detect misspecification of an SNM and decreases the likelihood that an SNM is non-identified. However, if the chosen  $\zeta$  is less than the true minimum time for which (15) holds bias will result. As a consequence, one should routinely include a table that shows how one’s estimate of  $E[Y_0]$  changes as  $\zeta$  is varied.

The RC and CD assumptions require that one record in  $X$  the minimum time of clinical onset among the set of clinical conditions whose preclinical phase could affect BMI. The exact clinical conditions that belong in this subset is a substantive question about which subject matter experts should be consulted. Also it is straightforward to generalize the CD assumption by replacing the constant  $\zeta$  with a known non-negative function  $\zeta\{\bar{L}(m)\}$  of past covariate history.

*Remark:* We will consider the effect of replacing the counterfactual  $X_m$  by the observed  $X$  in the CD assumption later.

*Estimation under a rank-preserving SNM for  $Y_m|X_m$  with  $X_m$  known.* To consistently estimate  $E[Y_0]$  under RC and CD, we must replace our SNMM model with an additive SNMM model for  $Y_m|X_m$  that also conditions on and allows effect modification by the counterfactual  $X_m$ . For pedagogic purposes, in this subsection we return to locally rank-preserving models. A locally rank-preserving SNM for  $Y_m|X_m$  states that

$$Y_{m+1} - Y_m = \gamma_m[A(m), \bar{A}(m-1), \bar{L}(m), X_m, \beta^*] \quad (16)$$

where  $\beta^*$  is an unknown parameter and  $\gamma_m[A(m), \bar{A}(m-1), \bar{L}(m), X_m, \beta]$  is a known function that can now depend on  $X_m$  that takes the value zero if either  $A(m)=0$  or  $\beta=0$ . (We emphasize that it is  $X_m$  and not  $\bar{X}(m)$  that occurs in the last equation.) This model is equivalent to assuming

$$Y_m = Y_m(\beta^*) \quad (17)$$

for each subject with  $Y_m(\beta)$  now redefined as

$$Y_m(\beta) = Y - \sum_{j=m}^K \gamma_m[A(j), \bar{A}(j-1), \bar{L}(j), X_j, \beta] \quad (18)$$

Now, of course the counterfactual variable  $X_m$  is itself unobserved. However for pedagogic purposes, in this subsection we unrealistically assume that in addition to the observed data  $(Y, X, \bar{L}, \bar{A})$ , data on the counterfactuals  $X_m$  are available.

*Remark:* We do not actually require a locally rank-preserving SNM for  $Y_m|X_m$ . A locally rank-preserving SNM for  $Y_m|c(X_m)$  for certain known functions  $c(x)$  could be used instead. This remark is explored further in the appendix.

Redefine  $Q_m(\beta) = q_m[\bar{L}(m), \bar{A}(m-1), X_m, Y_m(\beta)]$  to possibly be a function of  $X_m$ . Consider again the g-estimator  $\hat{\beta}$  that is equal to the  $\beta$  for which the 5 degree of freedom score test of  $\theta=0$  is precisely zero in the model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), X_m, \Xi(m) = 1, Y_m(\beta)] = \alpha^T W(m) + \theta^T Q_m(\beta)$$

$\hat{\beta}$  would be a consistent and asymptotically normal (CAN) estimator of  $\beta^*$  under the CO assumption, but not under the RC assumption. Under RC, the independence needed to make  $\theta=0$  when  $\beta=\beta^*$  only holds when we also condition on  $\bar{U}(m)$ .

However, consider the estimator  $\tilde{\beta}$  obtained when, for each time  $m$ , we only fit the previous model to subjects for whom  $X_m > m + \zeta$ , excluding all subjects with  $X_m \leq m + \zeta$ . This exclusion can be expressed by saying that we now fit the model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), X_m, \Xi(m) = 1, Y_m(\beta), X_m > m + \zeta] = \alpha^T W(m) + \theta^T Q_m(\beta) \quad (19)$$

Then the estimator  $\beta$  is the  $\beta$  for which the 5 degree of freedom score test of the hypothesis  $\theta=0$  is precisely zero in this latter model. When  $X_m > m + \zeta$ ,  $\bar{U}(m) = \bar{0}(m)$ , by assumption CD. Hence, we can rewrite the Equation (19) as

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), X_m, \Xi(m) = 1, Y_m(\beta), X_m > m + \zeta, \bar{U}(m) = \bar{0}(m)] = \alpha^T W(m) + \theta^T Q_m(\beta) \quad (20)$$

showing that we have succeeded in conditioning on  $\bar{U}(m) = \bar{0}(m)$ , even though  $\bar{U}(m)$  is unmeasured! It follows that, when the parameter  $\beta^*$  is identified, the estimator  $\tilde{\beta}$  is a CAN estimator of  $\beta^*$  under the RC and CD assumptions, as these assumptions imply the coefficient  $\theta=0$  if  $\beta=\beta^*$ . However as discussed further below, under the RC and CD assumptions, the positivity assumption no longer suffices to guarantee identification.

In summary, all that was required to produce a CAN estimator  $\tilde{\beta}$  of the parameter  $\beta^*$  of our locally rank-preserving SNM (17) for  $Y_m|X_m$  under the RC and CD assumptions was to restrict the earlier g-estimation procedure at each time  $m$  to those subjects with  $X_m > m + \zeta$ .

Thus, if  $\gamma_m[A(m), \bar{A}(m-1), \bar{L}(m), X_m, \beta] = A(m)\beta^T R_m$  is linear in  $\beta$  with  $R_m = r_m(\bar{L}(m), X_m)$  being a vector of known functions that can depend on  $X_m$ , then, given the OLS estimator  $\hat{\alpha}^T$  of  $\alpha^T$  in the model  $E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1] = \alpha^T W(m)$  and  $Q_m(\beta) = Q_m^* Y_m(\beta)$  linear in  $Y_m(\beta)$ , the CAN estimator  $\tilde{\beta}$  exists in closed form as

$$\tilde{\beta} = \left\{ \sum_{i=1, m=0}^{i=n, m=K} I[X_{im} > m + \zeta] \Xi_i(m) G_{im}(\hat{\alpha}) Q_{im}^* S_{im}^T \right\}^{-1} \times \left\{ \sum_{i=1, m=0}^{i=n, m=K} I[X_{im} > m + \zeta] \Xi_i(m) Y_i G_{im}(\hat{\alpha}) Q_{im}^* \right\} \quad (21)$$

with  $G_{im}(\hat{\alpha}) = A_i(m) - \hat{\alpha}^T W_i(m)$ ,  $S_{im} = \sum_{j=m}^K A_i(j) R_{ij}$ .

From the above, it follows that if, in addition to the observed data  $(Y, X, \bar{L}, \bar{A})$ , data on the counterfactuals  $X_m$  are available for each  $m$ , the sample average  $\sum_i^m Y_{0i}(\tilde{\beta})/n$  is a CAN estimator of the parameter of interest  $E[Y_0]$  under the RC and CD assumptions, provided  $\beta^*$  is identified and both our locally rank-preserving SNM for  $Y_m|X_m$  and our model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1] = \alpha^T W(m) \quad (22)$$

are correct. Of course data on  $X_m$  are unavailable. However in the next subsection, we prove an analog of this result holds without data on  $X_m$  under a locally rank-preserving SNFTM for the  $X_m$  which allows us to replace  $X_m$  by an estimate  $X_m(\tilde{\psi})$ , where  $\tilde{\psi}$  estimates the parameter  $\psi^*$  of our SNFTM.

Before proceeding to the next subsection, several additional points need to be made.

*Can we replace  $X_m$  by  $X$ :* A natural question that arises is the following. Suppose we replaced  $X_m$  by the observed  $X$  in the CD assumption, in our definition of  $Y_m(\beta)$ , and wherever else  $X_m$  occurs in this subsection, with the exception of the RC assumption (as the RC assumption with  $X$  replacing  $X_m$  would clearly be false if BMI is a cause of  $T$  and/or  $D$  and thus of  $X$ ). Do  $\tilde{\beta}$  and  $\sum_i^m Y_{0i}(\tilde{\beta})/n$  remain CAN estimators of  $\beta^*$  and  $E[Y_0]$ ? This question is natural in the sense that it is not obvious that the CD assumption and RP SNM based on  $X_m$  are more likely to be true than when based on  $X$ . So if the answer is ‘yes’ it would be simpler and more straightforward to use  $X$  in place of  $X_m$ . In particular, as  $X$ , unlike  $X_m$ , is observed, we would eliminate the need to replace  $X_m$  with the estimator  $X_m(\tilde{\psi})$ , thereby greatly simplifying the analysis.

Unfortunately,  $\tilde{\beta}$  and thus  $\sum_i^m Y_{0i}(\tilde{\beta})/n$  do not remain consistent when we use  $X$  in place of  $X_m$ . For this, consider the model

$$E[A(m)|\bar{L}(m), \Xi(m) = 1, \bar{A}(m-1), X, Y_m(\beta), X > m + \zeta] = \alpha^T W(m) + \theta^T Q_m Y_m(\beta) \quad (23)$$

which has replaced  $X_m$  in Equation (19) with  $X$ . Clearly,  $\tilde{\beta}$  will only be consistent for the parameter  $\beta^*$  of our locally RP SNM if  $\theta = 0$  when  $\beta = \beta^*$ . That is  $\tilde{\beta}$  will only be consistent if  $Y_m = Y_m(\beta^*)$  is independent of  $A(m)$  given  $(\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1, X, X > m + \zeta)$ . Now by the CD assumption with  $X$  replacing  $X_m$ ,  $X > m + \zeta$  implies  $\bar{U}(m) = \bar{0}(m)$ . Thus, consistency of  $\tilde{\beta}$  requires  $Y_m(\beta^*)$  independent of  $A(m)$  given  $(\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1, X, X > m + \zeta, \bar{U}(m) = \bar{0}(m))$ . However, we show in the next paragraph that this independence statement is not implied by the RC assumption and thus will generally be false, unless  $A(k)$  has no causal effect on  $X$  for  $k \geq m$  in which case  $X = X_m$  for each subject and we are back to Equation (20).

When  $A(m)$  has a causal effect on  $X$  (whether directly or through  $A(k)$ ,  $k > m$ ) then  $X$  is a common effect of two causes  $A(m)$  and  $X_m$  that are independent conditional on the event  $(\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1, \bar{U}(m) = \bar{0}(m), Y_m(\beta^*))$ . Therefore, conditional on both the previous event and  $(X, X > m + \zeta)$ ,

$A(m)$  and  $X_m$  are dependent and thus so are  $A(m)$  and  $Y_m(\beta^*)$ , as  $X_m$  and  $Y_m(\beta^*)$  are highly correlated, since both are functions of  $T_m$ .

However, even when  $A(m)$  has a causal effect on  $X$ , a slight modification of the above estimation procedure can be used to obtain CAN estimators of  $\beta^*$  in the special case in which  $A(m)$  has a known minimal latent period (MLP)  $\chi$  for its effect on  $X$  of at least  $\zeta$  months.

*Definition of MLP for effect on  $X$ :*  $A(m)$  has an MLP for its effect on  $X$  of  $\chi$  months if, for every subject and each time  $k > m$ ,  $X_k > m + \chi \Leftrightarrow X_m > m + \chi$  and  $X_k = X_m$  if  $X_m < m + \chi$ . In particular, by taking  $k = K + 1$ , the last two statements become  $X > m + \chi \Leftrightarrow X_m > m + \chi$  and  $X = X_m$  if  $X_m < m + \chi$ .

When a known MLP  $\chi$  exceeds  $\zeta$ , we can obtain CAN estimators of  $\beta^*$  by simply replacing  $X > m + \zeta$  by  $m + \chi > X > m + \zeta$  in model (23) as then the event  $(X, m + \chi > X > m + \zeta)$  is the event  $(X_m, m + \chi > X_m > m + \zeta)$  and we are back in the setting of Equation (19), except for the additional restriction,  $m + \chi > X_m$ , which does not introduce bias. Thus, the existence of an MLP of length  $\chi$  greater than  $\zeta$  allows us to estimate  $\beta^*$  and  $E[Y_0]$  without the need to specify an SNFTM for the  $X_m$ .

We now prove that under the RC and CD assumptions, an MLP of length  $\chi$  greater than  $\zeta$  implies that

$$X \prod [A(m)|\bar{L}(m), \Xi(m) = 1, \bar{A}(m-1), m + \chi > X > m + \zeta]$$

It follows that taking the RC and CD assumptions as given, we can test the hypothesis that an MLP of length  $\chi$  greater than  $\zeta$  exists by testing whether the last equation is true. To prove this claim note that, by the MLP assumption, the event  $m + \chi > X > m + \zeta$  is the event  $m + \chi > X_m > m + \zeta$  which, by the CD assumption, is the event  $m + \chi > X_m > m + \zeta, \bar{U}(m) = \bar{0}(m)$ . Thus, the equation in the last display is under the MLP and CD assumptions equivalent to the statement ‘ $X_m$  is independent of  $A(m)$  given  $(\bar{L}(m), \Xi(m) = 1, \bar{A}(m-1), m + \chi > X_m > m + \zeta, \bar{U}(m) = \bar{0}(m))$ ,’ which is true by the RC assumption.

Most experts believe it to be substantively implausible that an increase in BMI has a minimum latent period of more than 72 months, our default choice for  $\zeta$ . In contrast, in occupational cohort studies of the effect of a chemical carcinogen on time to clinical cancer, minimum latent periods of up to 10 years are commonly assumed.

*Estimation of  $E[Y_0]$  under a rank-preserving SNFTM.* As mentioned above, an analog of the above results hold when data on  $X_m$  are unavailable under a locally rank-preserving SNFTM for  $X_m$ . The simplest locally rank-preserving SNFTM specifies that

$$X_m = m + \int_m^X \exp(\psi^* A(t)) dt \quad \text{if } X > m \quad (24)$$

$$X_m = X \quad \text{if } X \leq m \quad (25)$$

where  $\psi^*$  is an unknown parameter and  $A(t)$  is as defined previously when  $t$  is a whole number of months and

$A(t) = A(\lfloor t \rfloor)$  when  $t$  is not a whole number where  $\lfloor t \rfloor$  is the largest integer less than or equal to  $t$ . Thus, by the definition of an integral as the area under a curve,

$$\int_m^X \exp(\psi^* A(t)) dt = \sum_{j=m}^{\lfloor X \rfloor} \exp(\psi^* A(j)) + \{X - \lfloor X \rfloor\} \exp(\psi^* A(\lfloor X \rfloor))$$

A locally rank-preserving SNFTM directly maps an individual's observed failure time  $X$  to the failure time  $X_m$  the individual would have under the 'time  $m$  dietary intervention.' Thus, it is a model for individual causal effects. If  $\psi^* = 0$ ,  $\exp(\psi^* A(t)) = 1$  and thus  $X_m = m + \int_m^X dt = m + X - m = X$  for any  $m$ . Hence  $\psi^* = 0$  encodes the sharp null hypothesis that  $X_0 = X$  for all subjects, that is, the 'time 0 dietary intervention' has no effect on any subject's  $X = \min(T, D)$ . It is useful to note that when  $\psi^* \neq 0$ , the SNFTM (24)–(25) implies that there is no minimal latent period for the effect of treatment on  $X$ . A general class (although not the most general class) of locally RP SNFTMs that includes the above one parameter model assumes

$$X_m = m + \int_m^X \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi^*)\} dt \quad \text{if } X > m \quad (26)$$

$$X_m = X \quad \text{if } X \leq m \quad (27)$$

where  $\omega(\bar{A}(t), \bar{L}(t), \psi) \equiv \omega(A(t), \bar{A}(t^-), \bar{L}(t), \psi)$ , is a known function satisfying  $\omega(A(t), \bar{A}(t^-), \bar{L}(t), \psi) = 0$  if  $A(t) = 0$  or  $\psi = 0$  and  $\bar{A}(t^-)$  is the A-history until just prior to time  $t$ . For example, we might have  $\omega(A(t), \bar{A}(t^-), \bar{L}(t), \psi) = A(t) \{\psi_0 + \psi_1^T \bar{L}(t)\}$  where  $L(t) = L(\lfloor t \rfloor)$  and  $L(\lfloor t \rfloor)$  is as defined earlier.

We next turn to estimation of  $\psi^*$ . For the moment, suppose the CO assumption modified to have  $(Y_m, X_m)$  in place of  $Y_m$  held and that  $(Y, X, \bar{L}, \bar{A})$  was observed. Then we could consistently estimate  $\psi^*$  by g-estimation. Specifically, we define

$$X_m(\psi) = m + \int_m^X \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt \quad \text{if } X > m \quad (28)$$

$$X_m(\psi) = X \quad \text{if } X \leq m \quad (29)$$

so under our model,  $X_m = X_m(\psi^*)$ . Note that, for each  $\psi$ ,  $X_m(\psi)$  can be computed from the observed data. Suppose, for concreteness,  $\psi^*$  is 5D so we search over a 5D grid. We let  $Q_m^{**}(\psi) = q_m^{**}[\bar{L}(m), \bar{A}(m-1), X_m(\psi)]$  be a 5D vector of functions of  $(\bar{L}(m), \bar{A}(m-1), X_m(\psi))$  such as  $Q_m^{**}(\psi) = X_m(\psi) [1, m, L^T(m)]$ . We use an extended linear model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1, X_m(\psi)] = \alpha^T W(m) + \theta^T Q_m^{**}(\psi)$$

Our g-estimate  $\hat{\psi}$  is the  $\psi$  for which the 5 degree of freedom score test that all five components of  $\theta$  equal zero is precisely zero. Since, by the modified CO assumption,  $\theta = 0$  if  $\psi = \psi^*$ , the g-estimate  $\hat{\psi}$  is CAN for  $\psi^*$ . The particular choice of

the functions  $Q_m^{**}(\psi)$  does not affect the consistency of the point estimate, but it determines the width of its confidence interval. Because  $X_m(\psi)$  is a nonlinear function of  $\psi$ , there is no closed form expression for  $\hat{\psi}$ . However, the equation solved by  $\hat{\psi}$  is a smooth function of  $\psi$ , so standard methods for solving nonlinear equations such as the Newton–Raphson algorithm can be used to compute  $\hat{\psi}$ .

Next suppose the observed data are still  $(Y, X, \bar{L}, \bar{A})$ , but the modified CO assumption does not hold. Rather, the CD and RC assumptions hold. Define the estimator  $\tilde{\psi}$  as the  $\psi$  for which the 5 degree of freedom score test of the hypothesis  $\theta = 0$  is precisely zero in the model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1, X_m(\psi), X_m(\psi) > m + \zeta] = \alpha^T W(m) + \theta^T Q_m^{**}(\psi)$$

Note the set of subjects who do not contribute to the score test of  $\theta = 0$  (that is, subjects with  $X_m(\psi) \leq m + \zeta$ ) depends on  $\psi$ . When  $X_m = X_m(\psi^*) > m + \zeta$ , then  $\bar{U}(m) = \bar{O}(m)$ , by assumption CD. Hence, at  $\psi = \psi^*$ , our procedure conditions on  $\bar{U}(m) = \bar{O}(m)$ . It follows that, provided  $\psi^*$  is identified, the estimator  $\tilde{\psi}$  is a CAN estimator of  $\psi^*$  under the RC assumption, as that assumption implies the coefficient  $\theta = 0$  if  $\psi = \psi^*$ . However, under the CD and RC assumptions, the positivity assumption does not guarantee identification.

Now let  $\beta(\tilde{\psi})$  be defined like  $\tilde{\beta}$  except that everywhere  $X_m(\tilde{\psi})$  replaces  $X_m$ , so that  $\tilde{\beta}(\tilde{\psi})$  is a function of the data  $(Y, X, \bar{L}, \bar{A})$  only. Next define

$$Y_m(\beta, \psi) = Y - \sum_{j=m}^K \gamma_m [A(j), \bar{A}(j-1), \bar{L}(j), X_j(\psi), \beta] \quad (30)$$

Note, by both models (17) and (26) being locally rank preserving,  $Y_m(\beta^*, \psi^*) = Y_m$ . Thus, when  $\psi^*$  and  $\beta^*$  are identified, the sample average  $\sum_i^n Y_{0i}[(\tilde{\psi}), \tilde{\psi}]/n$  is a CAN estimator of the parameter of interest  $E[Y_0]$  under the RC and CD assumptions, provided both our locally rank-preserving SNM (17) for  $Y_m|X_m$ , our locally rank-preserving SNFTM (26) for  $X_m$ , and our model (22) are all correctly specified.

*Estimation of  $E[Y_0]$  under an SNMM and an SNFTM without rank preservation.* As discussed earlier, the assumption of local rank preservation is biologically implausible. Thus, we will no longer assume that our locally rank-preserving models (17), and (26) are true. As a consequence, we can no longer assume that there exists some  $(\beta^*, \psi^*)$  such that the unobserved counterfactuals  $(X_m, Y_m)$  equal the observed  $(X_m(\psi), Y_m(\beta, \psi))$  when  $(\beta, \psi) = (\beta^*, \psi^*)$ . However, suppose with  $(X_m(\psi), Y_m(\beta, \psi))$  still defined by (28), (29), and (30), we assume that, for each  $m$ , there exists some  $(\beta^*, \psi^*)$  such that

*Assumption (i):* when  $\psi = \psi^*$ ,  $X_m$  and  $X_m(\psi)$  have the same conditional distribution given  $(A(m), \bar{L}(m), \bar{A}(m-1))$  and

*Assumption (ii):*

$$E[Y_m|A(m), \bar{L}(m), \bar{A}(m-1), X_m = x] = E[Y_m(\beta^*, \psi^*)|A(m), \bar{L}(m), \bar{A}(m-1), X_m(\psi^*) = x] \quad (31)$$

In contrast with the assumption of local RP, there is no *a priori* biological reason to exclude the possibility that both (i) and (ii) hold.

When assumptions (i) and (ii) hold for each  $m$ , we say the SNMM

$$\gamma_m[A(m), \bar{A}(m-1), \bar{L}(m), x, \beta] \tag{32}$$

for  $Y_m|X_m$  and the SNFTM (28)–(29) for  $X_m$  jointly hold with true parameter  $(\beta^*, \psi^*)$ . If the RC and CD assumptions, the model (22) and (i) and (ii) all hold, then  $\tilde{\psi}, \tilde{\beta}(\tilde{\psi})$  and  $\sum_i^m Y_{0i}[\tilde{\beta}(\tilde{\psi}), \tilde{\psi}]/n$  as defined previously are CAN for  $\psi^*, \beta^*$ , and the parameter of interest  $E[Y_0]$ , respectively, provided  $(\beta^*, \psi^*)$  are identified and we choose  $Q_m(\beta)$  linear in  $Y_m(\beta)$ . (In contrast,  $Q_m^*(\psi)$  need not be chosen linear in  $X_m(\psi)$ .) In summary,  $\tilde{\psi}, \tilde{\beta}(\tilde{\psi})$  and  $\sum_i^m Y_{0i}[\tilde{\beta}(\tilde{\psi}), \tilde{\psi}]/n$  have the same statistical properties under our joint SNMM model for  $Y_m|X_m$  and SNFTM for  $X_m$  when local rank preservation does not hold as when it does.

*Are remarkable results due to some sleight of hand.* The result summarized in the last sentence is striking for a number of reasons. Our comparability assumption, that is, the RC assumption, only assumes no unmeasured confounding conditional on  $\bar{U}(m)$ . Yet neither the SNMM for  $Y_m|X_m$  nor the SNFTM for  $X_m$  is a model for causal effects conditional on the unmeasured  $\bar{U}(m)$ . Thus, it is remarkable that these models can be used to estimate causal contrasts such as  $E[Y_0]-E[Y]$  under the RC and CD assumptions. Furthermore, even though  $X_m > m + \zeta$  implies  $\bar{U}(m) = \bar{0}(m)$  by the CD assumption, nonetheless, in the absence of local rank preservation,  $X_m(\psi^*) > m + \zeta$  does not imply  $\bar{U}(m) = \bar{0}(m)$ . Hence when local rank preservation does not hold, even though we condition on  $X_m(\psi) > m + \zeta$  in computing our g-estimates  $\tilde{\psi}, \tilde{\beta}(\tilde{\psi})$ , we do not thereby restrict the analysis to a subset of subjects all of whom have the same value of  $\bar{U}(m)$ ; thus one might guess confounding by the unmeasured  $\bar{U}(m)$  has not been controlled and our estimates of  $\tilde{\psi}, \tilde{\beta}(\tilde{\psi})$  and  $\sum_{i=1}^m Y_{0i}[\tilde{\beta}(\tilde{\psi}), \tilde{\psi}]/n$  must be inconsistent. Remarkably, such is not the case.

How did we pull off the seemingly remarkable ‘magic’ described in the preceding paragraph? We shall investigate whether we used some subtle ‘sleight of hand.’ We use a simple paradigmatic instance of our model that only involves a single time-independent exposure to guide our investigation. Specifically, we next provide an explicit proof that contains no ‘sleight of hand’ of our results in the case of a time-independent exposure. The general case is treated in the appendix. The reader who is interested more in the methodology and less interested in foundational issues may feel free to skip ahead to the section Censoring.

*Paradigmatic instance of a time-independent exposure:* We suppose that  $K + 1 = 1$  so time 0 is the only time of exposure. Further we assume there are no covariates. In this setting, the RC assumption becomes  $(Y_0, X_0)$  independent of  $A(0)$  given the unmeasured confounder  $U(0) = 0$ . The CD assumption becomes  $X_0 > \zeta$  implies  $U(0) = 0$ . Our SNFTM for  $X_0$  becomes

*Assumption (i):*  $X_0(\psi) = X \exp(\psi A(0))$  and  $X_0$  have the same conditional distribution given  $A(0)$  at  $\psi = \psi^*$ , whereas our SNMM for  $Y_m|X_m$  becomes

*Assumption (ii):*  $E[Y_0|A(0), X_0 = x] = E[Y_0(\beta^*, \psi^*)|A(0), X_0(\psi^*) = x]$  where  $Y_0(\beta, \psi) = Y - \gamma_0[A(0), X_0(\psi), \beta]$ .

Neither model makes any reference to  $U(0)$  and thus neither is a model for causal effects conditional on  $U(0)$ . Furthermore, although  $X_0 > \zeta$  implies  $U(0) = 0$  by the CD assumption, nonetheless  $X_0(\psi^*) > \zeta$  does not imply  $U(0) = 0$ . Now to prove our results.

*Proofs of our results:*

*Proof that  $\tilde{\psi}$  is CAN for  $\psi^*$ :* By assumption (i),  $\text{pr}[X_0(\psi^*) > t|A(0), X_0(\psi^*) > \zeta] = \text{pr}[X_0 > t|A(0), X_0 > \zeta]$ . But, by the CD and then the RC assumptions,  $\text{pr}[X_0 > t|A(0), X_0 > \zeta] = \text{pr}[X_0 > t|A(0), X_0 > \zeta, U(0) = 0] = \text{pr}[X_0 > t|X_0 > \zeta, U(0) = 0]$ . Hence,  $\text{pr}[X_0(\psi^*) > t|A(0), X_0(\psi^*) > \zeta]$  is not a function of  $A(0)$ . We conclude that  $A(0)$  and  $X_0(\psi^*)$  are independent given  $X_0(\psi^*) > \zeta$ . Thus  $E[A(0)|X_0(\psi^*), X_0(\psi^*) > \zeta] = \alpha + \theta X_0(\psi^*)$  has coefficient  $\theta = 0$  so, when  $\psi^*$  is identified, the  $\psi^*$  for which the score test of  $\theta = 0$  takes the value 0 is CAN for  $\psi^*$ .

*Proof that  $\tilde{\beta}(\tilde{\psi})$  is CAN for  $\beta^*$ :* By assumption (ii),  $E[Y_0(\beta^*, \psi^*)|A(0), X_0(\psi^*) = x, X_0(\psi^*) > \zeta] = E[Y_0|A(0), X_0 = x, X_0 > \zeta]$ .

But, by the CD and then the RC assumptions,  $E[Y_0|A(0), X_0 = x, X_0 > \zeta] = E[Y_0|A(0), X_0 = x, X_0 > \zeta, U(0) = 0] = E[Y_0|X_0 = x, X_0 > \zeta, U(0) = 0]$  is not a function of  $A(0)$ . Thus,  $0 = E[Y_0(\beta^*, \psi^*) \{A(0) - E[A(0)|X_0(\psi^*), X_0(\psi^*) > \zeta]\}]$ . Hence  $0 = E[Y_0(\beta^*, \psi^*) \{A(0) - E[A(0)|X_0(\psi^*) > \zeta]\}]$ . As a consequence, the  $\tilde{\beta}(\tilde{\psi})$  for which the score test of  $\theta = 0$  takes the value 0 in the model,  $E[A(0)|X_0(\tilde{\psi}) > \zeta, Y_0(\beta, \tilde{\psi})] = \alpha + \theta Y_0(\beta, \tilde{\psi})$ , is CAN for  $\beta^*$ , when  $\beta^*$  and  $\psi^*$  are identified.

*Proof that  $\sum_{i=1}^m Y_{0i}[\tilde{\beta}(\tilde{\psi}), \tilde{\psi}]/n$  is CAN for  $E[Y_0]$ :*

$$\begin{aligned} E[Y_0] &= \iint E[Y_0|A(0), X_0 = x] dF_{X_0}(x|A_0) dF(A_0) \\ &= \iint E[Y_0(\beta^*, \psi^*)|A(0), X_0(\psi^*) = x] \\ &\quad \times dF_{X_0(\psi^*)}(x|A_0) dF(A_0) = E[Y_0(\beta^*, \psi^*)] \end{aligned}$$

by assumptions (i) and (ii). Hence,  $\sum_{i=1}^m Y_{0i}[\tilde{\beta}(\tilde{\psi}), \tilde{\psi}]/n$  is CAN for  $E[Y_0(\beta^*, \psi^*)] = E[Y_0]$ , when  $\beta^*$  and  $\psi^*$  are identified.

This completes the promised proof of our results in the time-independent case. The proof in the appendix of the general time-dependent case is not much more difficult when one proceeds by induction. We conclude no sleight of hand occurred in the proof.

*Do correctly specified SNMMs for  $Y_m|X_m$  and SNFTMs for  $X_m$  always exist?* Perhaps the sleight of hand occurred right at the start, when we supposed that there exist  $(\beta^*, \psi^*)$  such that assumptions (i) and (ii) hold. We now prove that no such sleight of hand is afoot. Specifically, we prove that there always exist correctly specified SNMMs for  $Y_m|X_m$  and SNFTMs for  $X_m$ . (This result does not, of course, imply that the particular SNMM and SNFTM we actually choose to

analyze are correct.) We actually prove this result for an alternative, more intuitive, definition of an SNMM for  $Y_m|X_m$  and an SNFTM for  $X_m$  and then prove these alternative definitions are logically equivalent to assumptions (i) and (ii). This is done in this subsection for the special case of a time-independent exposure and in the appendix for a general time-varying exposure.

Consider again, for simplicity, our paradigmatic instance. Write  $A(0)$  as  $A$ . Suppose that  $Y, Y_0, X, X_0$  are all non-negative continuous random variables with support on  $(0, \infty)$ , satisfying the consistency assumption  $X = X_0$  and  $Y = Y_0$  if  $A = 0$ . Let  $S(x|A) = \text{pr}(X > x|A)$ . Let  $S_0(x|A) = \text{pr}(X_0 > x|A)$ . Let  $S_0^{-1}(x|A)$  be the inverse of  $S_0(x|A)$  with respect to the  $x$  argument. Define the function  $x_0^\dagger(x, A) = S_0^{-1}\{S(x|A)\}|A$ . Substituting 0 for  $A$ , we find  $x_0^\dagger(x, 0) = x$ , so

$$x_0^\dagger(X, 0) = X \text{ wp1} \quad (33)$$

Define  $X_0^\dagger = x_0^\dagger(X, A)$ . Then  $X_0^\dagger = x_0^\dagger(X, 0) = X$ , when  $A = 0$ . It is well known that  $X_0^\dagger = x_0^\dagger(X, A)$  and  $X_0$  have the same conditional distribution given  $A$ .

Define  $S(t|A, X_0^\dagger = x) = \text{pr}(Y > t|A, X_0^\dagger = x)$  and  $S_0(t|A, X_0 = x) = \text{pr}(Y_0 > t|A, X_0 = x)$ . Let  $S_0^{-1}(t|A, X_0 = x)$  be the inverse of  $S_0(t|A, X_0 = x)$  with respect to the  $t$  argument. Let  $y_0^\dagger(t, x, A) = S_0^{-1}\{S(t|A, X_0^\dagger = x)\}|A, X_0^\dagger = x$  and  $Y_0^\dagger = y_0^\dagger(Y, X, A)$ . Then  $Y_0^\dagger|A, X_0^\dagger = x$  and  $Y_0|A, X_0 = x$  have the same conditional distribution. It follows that  $(Y_0^\dagger, X_0^\dagger)|A$  and  $(Y_0, X_0)|A$  have the same joint conditional distribution. Thus,

$$E[Y_0^\dagger|X_0^\dagger = x, A] = E[Y_0|X_0 = x, A] \quad (34)$$

Define

$$\begin{aligned} \gamma^\dagger(A, x) &= E[Y|X_0^\dagger = x, A] - E[Y_0^\dagger|X_0^\dagger = x, A] \\ &\equiv E[Y - Y_0^\dagger|X_0^\dagger = x, A] \end{aligned} \quad (35)$$

The last two equations imply that

$$E[Y - \gamma^\dagger(A, X_0^\dagger)|X_0^\dagger = x, A] = E[Y_0|X_0 = x, A] \quad (36)$$

and

$$\gamma^\dagger(0, X) = 0 \text{ wp1} \quad (37)$$

as, by  $Y_0^\dagger = y_0^\dagger(Y, X, A)$ ,  $Y_0^\dagger = Y$  when  $A = 0$ .

Here are the alternative definitions of an SNFTM for  $X_0$  and an SNMM for  $Y_0|X_0$ .

**Definition a:** Let  $x_0(t, a, \psi)$  be a known function monotone increasing in  $t$  for each  $(a, \psi)$  satisfying  $x_0(t, a, \psi) = 1$  if  $a = 0$  or  $\psi = 0$ . We say  $x_0(t, a, \psi)$  is a correctly specified SNFTM for  $X_0$  if there exists  $\psi^*$  such that  $X_0(\psi^*) \equiv x_0(X, A, \psi^*)$  equals  $X_0^\dagger$  with probability one.

**Definition b:** We say a known function  $\gamma(a, x, \beta)$  satisfying  $\gamma(a, x, \beta) = 0$  if  $a = 0$  or  $\beta = 0$  is a correctly specified SNMM for  $Y_0|X_0$  if, for some  $\beta^*$ ,  $\gamma(A, X, \beta^*) = \gamma^\dagger(A, X)$  with probability one.

Define  $Y_0(\beta^*, \psi^*) = Y - \gamma(A, X_0(\psi^*), \beta^*)$ .

It is obvious from definitions (a) and (b) that there always exist correctly specified SNMMs for  $Y_0|X_0$  under definition (b) and correctly specified SNFTMs for  $X_0$  under definition

(a) as  $\gamma^\dagger(A, X)$  and  $x_0^\dagger(X, A)$  are well-defined functions of  $(F, F_0)$  satisfying  $\gamma^\dagger(0, X) = 0$  and  $x_0^\dagger(X, 0) = X$  with probability one, where  $F$  and  $F_0$ , respectively, denote the joint distribution of  $(Y, X, A)$  and of  $(Y_0, X_0, A)$ . Note  $\gamma^\dagger(A, X)$  and  $x_0^\dagger(X, A)$  do not depend on the conditional joint distribution of  $\{(Y, X), (Y_0, X_0)\}$  given  $A$ . This is as desired as this joint is not non-parametrically identified from data  $(Y, X, A)$  even when  $A$  is randomly assigned.

Thus, it only remains to show the logical equivalence of the original and alternative definitions of an SNFTM for  $X_0$  and an SNMM for  $Y_0|X_0$ .

The following Lemma shows that the alternative definitions of an SNFTM for  $X_0$  and an SNMM for  $Y_0|X_0$  imply the previous definitions.

**Lemma:** Suppose  $x_0(t, a, \psi)$  is a correctly specified SNFTM for  $X_0$  as defined in definition (a). Then  $X_0(\psi^*)|A$  has the same distribution as  $X_0|A$ . Further assume that  $\gamma(a, x, \beta)$  is a correctly specified SNMM for  $Y_0|X_0$  as defined in definition (b). Then  $E[Y_0(\beta^*, \psi^*)|A, X_0(\psi^*) = x] = E[Y_0|X_0 = x, A]$ .

**Proof:** The first result follows immediately from  $X_0^\dagger$  and  $X_0$  having the same conditional distribution given  $A$ . The second result follows from  $E[Y - \gamma^\dagger(A, x)|X_0^\dagger = x, A] = E[Y_0|X_0 = x, A]$ .

Finally, the following Lemma shows that the original definitions imply the alternative definitions.

**Lemma:** Suppose  $x_0(t, a, \psi)$  is monotone increasing in  $t$  for each  $(a, \psi)$  satisfying  $x_0(t, a, \psi) = 1$  if  $a = 0$  or  $\psi = 0$ . Further suppose that  $X_0(\psi^*)|A$  has the same distribution as  $X_0|A$  wp1 where  $X_0(\psi) = x_0(X, A, \psi)$ . Then  $X_0(\psi)$  is a correctly specified SNFTM for  $X_0$  under definition (a). In addition, suppose that  $\gamma(a, x, \beta^*)$  is a function satisfying  $\gamma(a, x, \beta) = 0$  if  $a = 0$  or  $\beta = 0$ . Suppose  $E[Y - \gamma(A, x, \beta^*)|X^\dagger = x, A = a] = E[Y_0|X_0 = x, A = a]$  for all  $(x, a)$  in a set of probability 1 under the law of  $(X_0, A)$ . Then,  $\gamma(a, x, \beta)$  is a correctly specified SNMM for  $Y_0|X_0$  under definition (b).

**Proof:** The proof of the first part follows from the well-known result that  $X_0^\dagger = x_0^\dagger(X, A)$  is the only function  $h(X, A)$  of  $(X, A)$  satisfying  $h(X, A)|A$  has the same distribution as  $X_0|A$  wp1 and  $h(x, 0) = x$ . The second part is proved by showing that  $\gamma^\dagger(a, x)$  is the unique function  $h(a, x)$  with  $h(0, x) = 0$  satisfying  $E[Y - h(A, x)|X^\dagger = x, A = a] = E[Y_0|X_0 = x, A = a]$  for all  $(x, a)$  in a set of probability 1 as in Robins *et al.*<sup>11</sup> and Lok *et al.*<sup>12</sup>

Are  $\gamma^\dagger(a, x), x_0^\dagger(x, a)$  and  $E[Y_0]$  non-parametrically identified from data  $(Y, X, A)$  under our assumptions? In this subsection, we finally uncover some sleight of hand that provided us with such seemingly magical results. Although we restrict our discussion to the special case of a time-independent exposure, similar results apply in the general case. Specifically, we will show that  $\gamma^\dagger(a, x)$ ,  $x_0^\dagger(x, a)$  and  $E[Y_0]$  are not identified by the distribution of  $(Y, X, A)$  under the RC and CD assumptions. Previously, we saw that  $\gamma^\dagger(a, x)$ ,  $x_0^\dagger(x, a)$  and  $E[Y_0]$  are identified and equal  $\gamma(a, x, \beta^*)$ ,  $x_0(x, a, \psi^*)$  and  $E[Y - \gamma(A, X, \beta^*)]$ , respectively, when we assume a correctly specified SNFTM  $x_0(x, a, \psi)$  for  $X_0$  and an SNMM  $\gamma(a, x, \beta)$  for

$Y_0|X_0$  whose true parameters  $\psi^*$  and  $\beta^*$  are identified (by g-estimation). It follows that identification of  $\gamma^\dagger(a, x)$ ,  $x_0^\dagger(x, a)$  and  $E[Y_0]$  must result from the functional form restrictions encoded in our models  $x_0(x, a, \psi)$  and  $\gamma(a, x, \beta)$ . It follows that if we make the restrictions imposed by our models less rigid by adding additional parameters, we can lose identification of  $\gamma^\dagger(a, x)$ ,  $x_0^\dagger(x, a)$  and  $E[Y_0]$ . This loss of identification occurs when, in an infinite sample size, more than one combination of parameters, say the true parameters  $(\psi^*, \beta^*)$  and the false parameters  $(\psi^{**}, \beta^{**})$ , both make the score tests in our g-estimation procedures exactly zero for all choices of  $Q_m(\beta)$  linear in  $Y_m(\beta)$  and all choices of  $Q_m^{**}(\beta)$ . This loss of identification can be expressed by saying that the data (even were the sample size is infinite) cannot be used to determine whether the true causal quantities are  $\gamma(a, x, \beta^*)$ ,  $x_0(x, a, \psi^*)$  and  $E[Y_{-\gamma}(A, X, \beta^*)]$  versus  $\gamma(a, x, \beta^{**})$ ,  $x_0(x, a, \psi^{**})$  and  $E[Y_{-\gamma}(A, X, \beta^{**})]$ .

In contrast,  $\gamma^\dagger(a, x)$ ,  $x_0^\dagger(x, a)$  and  $E[Y_0]$  are identified under the comparability assumption that  $(Y_0, X_0)$  is independent of  $A_0$ , without any reliance on the functional form restrictions encoded in our models. However, in contrast with assumption RC, this comparability assumption contradicts our substantive knowledge, as it implies no unmeasured confounding by undiagnosed chronic disease.

The problem of lack of identification under the RC and CD assumptions has little to do with the question of local rank preservation. Suppose we have assumed a correctly specified SNFTM  $x_0(x, a, \psi)$  for  $X_0$  and we do not assume RP. Suppose in truth RP holds. Nonetheless, a second investigator who assumes the RP version of the SNFTM model gains nothing thereby with regard to the estimation of  $x_0^\dagger(x, a)$ : the causal quantity  $x_0^\dagger(x, a)$  is identified under the non-rank-preserving SNFTM if and only if it is identified under the RP SNFTM. However, a small amount could be gained by assuming rank preservation for an SNMM; rarely by assuming RP a non-identifiable SNMM can become identifiable as one can then use nonlinear functions  $Q_m(\beta)$  of  $Y_m(\beta)$  in g-estimation. But this advantage is not actually due to rank preservation. Rather it is due to the fact that an RP SNMM is actually a special case of a structural nested distribution model as defined in Robins<sup>9</sup> and Robins and Wasserman.<sup>3</sup> Our SNMM model  $\gamma(a, x, \beta)$  for  $Y_0|X_0$  is a structural nested distribution model if  $Y_{-\gamma}(A, X, \beta^*)$  is independent (rather than just mean independent) of  $A$  given  $X$ . It is this independence (rather than rank preservation) that licenses the use of nonlinear functions  $Q_m(\beta)$  of  $Y_m(\beta)$  in g-estimation.

*Non-identifiability of  $\gamma^\dagger(a, x)$ ,  $x_0^\dagger(x, a)$  and  $E[Y_0]$ :* Suppose we do not impose an SNFTM for  $X_0$  or an SNMM for  $Y_0|X_0$ . Then, it is clear that all we can conclude under assumptions RC and CD is that  $X_0^\dagger = x_0^\dagger(X, A)$  and  $A \equiv A(0)$  are independent given  $X_0^\dagger > \zeta$  and  $E[Y_{-\gamma_0^\dagger}(A, x)|A(0), X_0^\dagger = x, X_0^\dagger > \zeta] = E[Y_{-\gamma_0^\dagger}(A, x)|X_0^\dagger = x, X_0^\dagger > \zeta]$ . As a consequence, our parameter of interest  $E[Y_0]$  is not identified. Specifically, under RC and CD, with  $p = \text{pr}(A = 0)$

$$E[Y_0] = \tag{38}$$

$$E[Y|X > \zeta, A = 0] \{ \text{pr}[X > \zeta|A = 0]p + \{1 - \text{pr}[X^\dagger < \zeta|A \neq 0]\}(1 - p) \} \tag{39}$$

$$+ E[Y|X \leq \zeta, A = 0] \text{pr}[X \leq \zeta|A = 0]p \tag{40}$$

$$+ E[\{Y - \gamma^\dagger(A, X^\dagger)\}|X^\dagger \leq \zeta, A \neq 0] \text{pr}[X^\dagger \leq \zeta|A \neq 0] (1 - p) \tag{41}$$

However, the quantities

$$\text{pr}[X^\dagger \leq \zeta|A \neq 0] = \text{pr}[X_0 \leq \zeta|A \neq 0] \tag{42}$$

$$E[\{Y - \gamma^\dagger(A, X^\dagger)\}|X^\dagger \leq \zeta, A \neq 0] = E[Y_0|X_0 \leq \zeta, A \neq 0] \tag{43}$$

are not identified under the RC and CD assumptions. It suffices to show this when RP holds. So, for the moment assume RP. Because both quantities (42) and (43) refer to the distribution of the counterfactuals responses  $(Y_0, X_0)$  under no exposure (no weight gain) among those who actually were exposed ( $A \neq 0$ ), we need an assumption to identify these quantities under RP. But under RC, we only have comparability conditional on  $U(0) = 0$ , so identification fails.

When we additionally assume an SNFTM for  $X_0$  and an SNMM for  $Y_0|X_0$ , we may or may not obtain identification of  $E[Y_0]$  depending on whether the additional functional form restrictions encoded in the models suffice to identify the quantities (42) and (43) by allowing us to extrapolate from  $X_0 > \zeta$  where we have comparability (as, by CD,  $U(0) = 0$ ) to  $X \leq \zeta$  where we do not. To clarify this last statement, consider the following RP SNM for  $Y_0|X_0$ :  $Y_0 = Y_{-\gamma}(A, X_0, \beta^*)$  with

$$\gamma(A, X_0, \beta) = \beta_0 AI(X_0 \leq \zeta) + \beta_1 AI(X_0 > \zeta) \tag{44}$$

Under assumptions RC and CD, even if we unrealistically suppose that data on  $X_0$  was available for all subjects, we could not identify  $\beta^* = (\beta_0^*, \beta_1^*)^T$  because  $\beta_0^*$  would not be identified, although  $\beta_1^*$  would be identified. This follows from the fact that, under RC and CD, no subject with  $X_0 \leq \zeta$  may contribute to g-estimation of  $\beta^*$ . As a consequence, we cannot identify  $E[Y_0]$  because  $Y_0 = Y - \beta_0^*$  is not estimable on the subset of exposed subjects ( $A = 1$ ) with  $X_0 \leq \zeta$ .

In contrast, when data on  $X_0$  are available,  $\beta^*$  and  $E[Y_0]$  are identified in the RP SNM  $\gamma(A, X_0, \beta) = \beta_0 A + \beta_1 AX_0$  because both  $\beta_0^*$  and  $\beta_1^*$  can be estimated by g-estimation restricted to subjects with  $X_0 > \zeta$ . Thus  $Y_0 = Y - \beta_0^* A - \beta_1^* AX_0$  can be estimated for all subjects, including those with  $A = 1$  and  $X_0 \leq \zeta$ , because, by having the same parameters apply to subjects with  $X_0 \leq \zeta$  as to subjects with  $X_0 > \zeta$ , the model allows extrapolation from subjects with  $X_0 > \zeta$  to subjects with  $X_0 \leq \zeta$ . One must weigh the benefit of extrapolation that comes with assuming the model  $\gamma(A, X_0, \beta) = \beta_0 A + \beta_1 AX_0$  against the risk that the model is misspecified for subjects with  $X_0 \leq \zeta$ , as would be the case were the true model:  $\gamma(A, X_0, \beta^*) = \beta_0^* AI(X_0 > \zeta) + \beta_1^* AX_0 I(X_0 > \zeta) + \beta_2^* AI(X_0 \leq \zeta) + \beta_3^* AX_0 I(X_0 \leq \zeta)$  with  $\beta_2^*$  very different from  $\beta_0^*$  and with  $\beta_3^*$  very different from  $\beta_1^*$ . Then the extrapolated value  $Y - \beta_0^* A - \beta_1^* AX_0$  for  $Y_0$  based on the misspecified model would be a badly biased estimate of the true  $Y_0$  for subjects with  $A = 1$  and  $X_0 \leq \zeta$ . Yet because the

model  $\gamma(A, X_0, \beta) = \beta_0 A + \beta_1 A X_0$  is correct for subjects with  $X_0 > \zeta$ , there exists no valid test of model fit that could detect the biased extrapolation when we only assume RC and CD.

Suppose now, as is true in practice, data on  $X_0$  are unavailable for subjects with  $A = 1$ . Then, under assumptions RC and CD, without the help of a correct RP SNFTM for  $X_0$  whose functional form provides for extrapolation, we can no longer identify any aspect of the distribution of  $Y_0$  for any identifiable subset of subjects with  $A \neq 0$ . This is because, although we know that the identified quantity  $E[Y|X > \zeta, A = 0]$  equals  $E[Y_0|X_0 > \zeta, A \neq 0]$ , we cannot identify which subjects with  $A \neq 0$  have  $X_0 > \zeta$ .

In summary, in the realistic setting of longitudinal time-dependent exposures, the possibility of sensitivity of one's estimate of  $E[Y_0]$  to model extrapolation should be examined by reestimating  $E[Y_0]$  under a variety of models that differ in both the dimension of the parameter vectors and in functional form.

A final point is that no individual who has developed a chronic disease by time  $m$  is included in our g-estimation procedure at  $m$  because  $X_m(\psi) = X < m + \zeta$  for such subjects. Thus, our estimate of the effect of exposure at time  $m$  on a subject with a chronic disease at  $m$  is identified wholly by extrapolation from the effect on subjects without chronic disease at  $m$ . One approach to lessening the degree of extrapolation is to require a subject to be rather ill before they meet the definition of having a diagnosed chronic disease. For example, mild to moderate Db or hypertension need not qualify as having a chronic disease, especially if regular data on BP and blood glucose have been recorded in the database, as unmeasured confounding by undiagnosed mild to moderate Db or hypertension should then be minimal. If our definition of a diagnosed chronic disease is sufficiently stringent, then few subjects who meet the definition at  $m$  will be observed to gain weight subsequent to  $m$ . In that case, model-based extrapolation must be minimal—any model-based extrapolation is restricted to those gaining weight at  $m$ , because our models are models for the causal effect of weight gain (not loss) at  $m$ . In the section Intractable confounding in subgroups, we offer a different approach to lessening our reliance on model misspecification.

*Can we replace  $X_m$  by  $X$  revisited.* We revisit the issue of whether we could have replaced  $X_m$  by the observed  $X$  in the CD assumption if we are willing to assume an SNFTM for  $X_m$  so as to link the distribution of  $X$  with that of  $X_m$ . We take the observed data to be  $(\bar{A}(K), \bar{L}(K+1), Y, X)$ . We will study the implications of two different SNFTMs. The first SNFTM is the model discussed above that assumes  $X_m(\psi^*)$  and  $X_m$  have the same conditional distribution given  $(\bar{L}(m), \bar{A}(m))$ . The second assumes  $X_m(\psi^*)$  and  $X_m$  have the same conditional distribution given  $(\bar{L}(m), \bar{A}(m), \bar{U}(m) = 0)$ . In both cases,  $X_m(\psi^*)$  is defined by Equations (28) and (29). Note a locally RP SNFTM implies  $X_m(\psi^*) = X_m$  and thus

both models are true. When rank preservation does not hold, the truth of one model does not imply the truth of the other. We first show that when rank preservation does not hold, under the RC assumption and the modified CD assumption in which  $X_m$  is replaced by the observed  $X$ , the parameter  $\psi^*$  of the first SNFTM may not be identifiable; however, the parameter of the second model is estimable by g-estimation. Thus, one might assume we might impose the modified CD assumption and the second model in lieu of the unmodified CD assumption and the first model. However, we shall see this approach has a drawback: knowledge of the parameter  $\psi^*$  of the second model in contrast to that of the first model does not help identify the parameter of interest  $E[Y_0]$ .

We now show that  $\psi^*$  is identifiable in the second SNFTM model under RC and the modified CD assumptions. Note  $X > \zeta + m$  is equivalent to

$$X_m(\psi) = m + \int_m^X \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt > m + \int_m^{\zeta+m} \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt$$

Thus, the modified CD assumption implies that whenever

$$X_m(\psi) \geq m + \int_m^{\zeta+m} \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt$$

we have  $\bar{U}(m) = 0$ . However, even if we made the rank preservation assumption that  $X_m(\psi^*) = X_m$ , we cannot therefore conclude from the RC assumption that  $A(m)$  is independent of  $X_m(\psi^*)$  given  $(\bar{L}(m), \bar{A}(m), X_m(\psi^*) \geq m + \int_m^{\zeta+m} \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt)$ ; although this conditioning event indeed implies  $\bar{U}(m) = 0$ ; nonetheless, the conditioning event also depends on  $A(t)$  for  $t > m$ , whereas the conditioning events in the RC assumption do not.

However, if we let  $d(m, \psi, \zeta)$  be the maximum value of  $X_m(\psi)$  among all subjects with  $m < X < \zeta + m$  (that is, subjects with  $m < X_m(\psi) < m + \int_m^{\zeta+m} \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt$ ), then  $X_m(\psi) > d(m, \psi, \zeta)$  implies  $X > \zeta + m$  and thus  $\bar{U}(m) = 0$ . Thus, we can conclude from the RC assumption that, under a rank-preserving model,  $A(m)$  and  $X_m(\psi^*)$  are independent given  $(\bar{L}(m), \bar{A}(m), X_m(\psi^*) \geq d(m, \psi^*, \zeta))$ , as  $d(m, \psi, \zeta)$  does not vary among the subjects. (Technically, this independence only holds if we replace  $d(m, \psi, \zeta)$  by its probability limit. But this distinction is unimportant for inference because  $d(m, \psi, \zeta)$  converges to its probability limit at a rate even faster than  $n^{1/2}$  under mild regularity conditions.) Thus, given a rank-preserving SNFTM, we can use g-estimation to obtain a CAN estimate  $\tilde{\psi}$  of  $\psi^*$  under the RC and modified CD assumptions. Specifically,  $\tilde{\psi}$  is the  $\psi$  for which the 5 degree of freedom score test of the hypothesis  $\theta = 0$  is precisely zero in the model

$$E[A(m)|\bar{L}(m), \bar{A}(m-1), \Xi(m) = 1, X_m(\psi), X_m(\psi) > d(m, \psi, \zeta)] = \alpha^T W(m) + \theta^T Q_m^{**}(\psi)$$

Suppose now rank preservation is absent. If we assume the second SNFTM, we know  $X_m(\psi^*)$  and  $X_m$  have the same distribution given  $\bar{L}(m)$ ,  $\bar{A}(m)$ ,  $\bar{U}(m)=0$ . Thus, by the RC assumption  $A(m)$  and  $X_m(\psi^*)$  are independent given  $(\bar{L}(m), \bar{A}(m), X_m(\psi^*) \geq d(m, \psi^*, \zeta), \bar{U}(m)=0)$ . Hence,  $A(m)$  and  $X_m(\psi^*)$  are independent given  $(\bar{L}(m), \bar{A}(m), X_m(\psi^*) \geq d(m, \psi^*, \zeta))$  as the event  $\bar{L}(m), \bar{A}(m), X_m(\psi^*) > d(m, \psi^*, \zeta)$  is equivalent to the event  $\bar{L}(m), \bar{A}(m), X_m(\psi^*) > d(m, \psi^*, \zeta), \bar{U}(m)=0$ . So  $\tilde{\psi}$  generally remains CAN for  $\psi^*$ .

We next show that  $\psi^*$  is not identifiable in the first SNFTM model under RC and the modified CD assumptions. Under the first SNFTM, we only know  $X_m(\psi^*)$  and  $X_m$  have the same distribution given  $\bar{L}(m), \bar{A}(m)$ . Thus  $X_m(\psi^*) | \bar{L}(m), \bar{A}(m), X_m(\psi^*) > d(m, \psi^*, \zeta)$  has the same distribution as  $X_m | \bar{L}(m), \bar{A}(m), X_m > d(m, \psi^*, \zeta)$ .

Thus, by equivalence of the conditioning events, both  $X_m(\psi^*) | \bar{L}(m), \bar{A}(m), X_m(\psi^*) > d(m, \psi^*, \zeta)$  and  $X_m(\psi^*) | \bar{L}(m), \bar{A}(m), X_m(\psi^*) > d(m, \psi^*, \zeta), \bar{U}(m)=0$  have the same distribution as  $X_m | \bar{L}(m), \bar{A}(m), X_m > d(m, \psi^*, \zeta)$ . However, under the first SNFTM and without rank preservation, this equality does not allow us to invoke the RC assumption, as the conditioning event  $\bar{L}(m), \bar{A}(m), X_m > d(m, \psi^*, \zeta), \bar{U}(m)=0$  in that assumption differs from the conditioning event  $\bar{L}(m), \bar{A}(m), X_m > d(m, \psi^*, \zeta)$ . Thus, we cannot conclude  $A(m)$  and  $X_m(\psi^*)$  are independent given  $(\bar{L}(m), \bar{A}(m), X_m(\psi^*) \geq d(m, \psi^*, \zeta))$  and so  $\tilde{\psi}$  will not be CAN for  $\psi^*$  under the first SNFTM. Indeed identification is not possible.

Finally, we argue that knowledge of the parameter  $\psi^*$  of the second model in contrast to that of the first model does not help identify  $E[Y_0]$ . Under the second model, we only learn the causal effect of treatment at time  $m$  among those with  $\bar{U}(m)=0$ . This does not allow us to estimate the distributions of  $X_m$  and thus  $Y_m$  for all subjects. In fact, the counterfactual distribution of  $X_m$  and thus  $Y_m$  are not even identified in those with  $\bar{U}(m)=0$  for  $m < K$ , because the distributions of  $X_K$  and thus  $Y_K$  are not identifiable in those with  $\bar{U}(m)=0$  but  $\bar{U}(K) \neq 0$ . One way to understand the difference is that the second model does not allow for the extensive model-based extrapolation that the first model does. Whether that is viewed as a drawback of the second model clearly depends on one's faith in versus skepticism about model-based extrapolation.

#### Intractable confounding in subgroups

Our comparability assumption RC that  $A(m)$  is statistically independent of  $(Y_m, X_m)$  given both  $\bar{L}(m)$  and  $\bar{U}(m) = \bar{O}(m)$  at time  $m$  may not be reasonable for particular, identifiable subgroups of the study population. That is, there may be identifiable subgroups in whom confounding by unmeasured factors is intractable, where, by definition, a subgroup is identifiable at time  $m$  if membership in the subgroup is determined by the measured variables  $\bar{L}(m)$ . In the section Intractable unmeasured confounding in subgroups, we noted that possible examples of such subgroups include

subjects with a diagnosed chronic disease, an age of greater than 70 years, or a BMI below 21 kg/m<sup>2</sup>. In fact, as we have assumed  $U(m)=1$  whenever  $X < m$ , we have all along been assuming intractable confounding in the identifiable subgroup consisting of those alive with a diagnosed chronic disease at  $m$  ( $X < m, T > m$ ). We have therefore been excluding them from our g-estimation procedure by requiring  $X_m > m + \zeta$  for inclusion. Recall that if  $X < m$ , then  $X = X_m$ .

Suppose therefore we wish to conduct an analysis where no comparability assumption (neither CO nor RC) is assumed at time  $m$  for subjects who, at  $m$ , have an age of greater than 70 years, or a BMI below 21 kg/m<sup>2</sup>. To do so, we simply redefine  $\Xi(m)$  to be zero for such subjects regardless of whether or not their BMI( $m+1$ )  $\geq$  BMI<sub>max</sub>( $m$ ), so that they too are excluded from contributing to g-estimation at time  $m$ . In so doing, we do not change the models that are fit, the interventions under consideration or the parameter of interest  $E[Y_0]$ . Rather we only change, by decreasing, the number of person-time observations used to estimate our model parameters. We thereby sacrifice some power and efficiency. As a consequence, even where willing to make assumption CO for the remaining subjects with  $\Xi(m)=1$ ,  $E[Y_0]$  would no longer be non-parametrically identified, because model-based extrapolation is now being used for identification.

In contrast to g-estimation of SNMs, when confounding by unmeasured factors is present in certain subgroups of the study population, neither IPTW estimation nor the parametric g-formula estimator can be used to estimate  $E[Y_0]$ .

If a substantial fraction of the total person time is accrued by subjects in identifiable subgroups with intractable confounding, then either identification will fail or, more often, the validity of one's estimate of  $E[Y_0]$  will rely heavily on model extrapolation. One, albeit not altogether satisfactory, way to decrease the reliance on model extrapolation is to give up the attempt to estimate the parameter of interest  $E[Y_0]$ . Instead, let  $IN(m)$  be the indicator of intractable confounding in identifiable subgroups that takes the value 1 if at time  $m$  a subject is in an identifiable subgroup with intractable confounding and 0 otherwise. Note that, based on the above discussion, subjects alive at  $m$  with  $X < m$  have  $IN(m)=1$ .

Define  $Y_m^T$  to be one's counterfactual outcome when following the time  $m^T$  dietary intervention in which a subject follows his observed diet up through month  $m$  and is thereafter weighed daily. On any day in month  $k > m$  that his weight exceeds his previous maximum monthly weight, the subject's caloric intake is restricted whenever  $IN(k)=0$ . However, during months in which a subject is in an intractable subgroup [ $IN(k)=1$ ], we place no restrictions on his diet or weight gain, reflecting the fact that due to intractable confounding, we are unable to estimate the effect of preventing weight gain among subjects with  $IN(m)=1$ , except by model extrapolation. In this section a superscript T does not denote matrix transposition.

Our new goal becomes to estimate  $E[Y_0^T]$ , the mean utility under an intervention in which, starting at the age of 18 years, each time  $m$  a subject with  $IN(m)=0$  exceeds his past maximum BMI, we calorie restrict him to prevent further weight gain. To estimate  $E[Y_0^T]$  by g-estimation, we proceed exactly as above except (i) we define new variables  $A^T(m)$  and  $\Xi^T(m)$  that equal  $A(m)$  and  $\Xi(m)$  whenever  $IN(m)=0$  but are zero whenever  $IN(m)=1$ , and (ii) everywhere replace  $A(m)$  and  $\Xi(m)$  in our g-estimation procedure by  $A^T(m)$  and  $\Xi^T(m)$ . Then, our algorithm that had estimated  $E[Y_0^T]$  will now output an estimator of  $E[Y_0^T]$ . In summary, at the cost of estimating a parameter  $E[Y_0^T]$  of lesser interest than  $E[Y_0]$ , we have eliminated the model extrapolation required to estimate the effect of weight gain among subjects with  $IN(m)=1$ .

However, the procedure in the preceding paragraph has not eliminated the model extrapolation required to estimate the effect of weight gain among the intractably confounded non-identifiable subgroup defined by  $m < X_m < m + \varsigma$ . As a consequence  $E[Y_0^T]$ , like  $E[Y_0]$ , fails to be non-parametrically identified and must rely on model extrapolation for identification. Specifically, the subgroup with  $m < X_m < m + \varsigma$  is intractably confounded by  $U(m)$ . It is not identifiable because the observed data cannot determine membership. For example, among subjects with  $A^T(m) > 0$ , we cannot determine if a subject with  $X$  observed to be between  $m$  and  $m + \varsigma$  is a subject with  $m < X_m < m + \varsigma$  versus a subject with  $X_m < m + \varsigma$ , with  $X$  occurring before  $m + \varsigma$  owing to the causal effect of his weight gain  $A^T(m)$ . As a consequence, it is not possible to assign all members of the intractably confounded subgroup with  $m < X_m < m + \varsigma$  the value  $IN(m)=1$ , while assigning all members of the unconfounded subgroup with  $m + \varsigma < X_m$  the value  $IN(m)=0$ . The latter subgroup is unconfounded under the RC assumption because  $m + \varsigma < X_m$  implies  $\bar{U}(m) = \bar{0}(m)$  by the CD assumption.

In fact, an MLP with length  $\chi > \varsigma$  must exist for non-parametric identification of  $E[Y_0^T]$ . For the remainder of this subsection, assume such an MLP. Then subjects with  $m < X_m < m + \varsigma$  form an identifiable subgroup, as  $m < X < m + \varsigma$  and  $m < X_m < m + \varsigma$  are equivalent. Similarly, subjects with  $X_m > m + \varsigma$  now form an identifiable subgroup. Thus we can now assign  $IN(m)=1$  to all subjects in the confounded subgroup  $m < X_m < m + \varsigma$  and  $IN(m)=0$  to all members of the subgroup  $X_m > m + \varsigma$  who were not already known to have  $IN(m)=1$  by virtue of membership in some other intractably confounded subgroup (for example, age greater than 70 years). Once we have assigned all members of the subgroup  $m < X_m < m + \varsigma$  the value  $IN(m)=1$ , our time  $m^T$  dietary interventions no longer restrict the diet of any subject of any intractably confounded subgroup. As a consequence,  $E[Y_0^T]$  is now non-parametrically identified. A formal proof is given in the appendix where it is also shown that, owing to the non-parametric identification,  $E[Y_0^T]$  can be estimated using the parametric g-formula estimator and the IPTW estimator, as well as by g-estimation of SNMs.

## Censoring

We now consider the realistic setting in which the available data are  $O = \bar{A}(K), \bar{L}(K+1), Y, XI(X \leq K+1)$ , indicating that  $X$  is not observed in subjects for whom  $X$  exceeds the end of follow-up time  $K+1$ . For such censored subjects,  $X_m(\psi)$  is not observed. As a consequence, g-estimation as described above cannot be done. We will describe a modified estimation procedure that can be validly applied to censored data. In the interest of brevity, we only consider a procedure that is easy to describe. The down side is that the procedure we describe is not as efficient as other more complex procedures.

Given an SNFTM for  $X_0$  we can still use g-estimation to obtain CAN estimates  $\tilde{\psi}$  of  $\psi^*$  from censored data by replacing everywhere  $X_m(\psi)$  by  $C_m(\psi) = \min(X_m(\psi), K_m(\psi))$ , in the g-estimation procedure, where

$$K_m(\psi) = m + \min_{\{t: X_t > K+1\}} \left\{ \int_m^{K+1} \exp\{\omega(\bar{A}_i(t), \bar{L}_i(t), \psi)\} dt \right\} \quad (45)$$

is the smallest possible value of  $X_m(\psi)$  any censored subject could possibly have (as  $m + \{\int_m^{K+1} \exp\{\omega(\bar{A}(t), \bar{L}(t), \psi)\} dt\}$  would be  $X_m(\psi)$  for a given censored subject had he died, unbeknownst to us, immediately after end of follow-up). Note  $C_m(\psi) > m + \varsigma$  implies  $X_m(\psi) > m + \varsigma$  so our g-estimation procedures remain restricted to subjects with  $\bar{U}(m) = 0$ .

Similarly, given an SNMM model we can still use g-estimation to obtain CAN estimates  $\tilde{\beta}(\tilde{\psi})$  of  $\psi^*$  from censored data by replacing  $X_m(\psi)$  by  $C_m(\psi)$ , everywhere in the g-estimation procedure. However, there is a subtlety in interpretation. Specifically, define the function  $c_m(x, \psi) = \min(x, K_m(\psi))$ , so  $c_m(X_m(\psi), \psi) = C_m(\psi)$ . Set  $C_m = c_m(X_m, \psi^*)$ . The correct definition of our SNMM model is

$$E[Y_m|A(m), \bar{L}(m), \bar{A}(m-1), C_m = x] \quad (46)$$

$$= E[Y_m(\beta^*, \psi^*)|A(m), \bar{L}(m), \bar{A}(m-1), C_m(\psi^*) = x] \quad (47)$$

where now,

$$Y_m(\beta, \psi) = Y - \sum_{j=m}^K \gamma_j[A(j), \bar{A}(j-1), \bar{L}(j), C_j(\psi), \beta] \quad (48)$$

We refer to this model as an SNMM model for  $Y_m|C_m$ . Technical details are given in the appendix. Finally, a CAN estimator of  $E[Y_0]$  from censored data is  $\sum_{i=1}^n Y_{0,i}[(\tilde{\beta}(\tilde{\psi}), \tilde{\psi})]/n$  as before with  $\tilde{\beta}(\tilde{\psi})$  and  $\tilde{\psi}$  as redefined in this section.

## Maximum weight gain dietary intervention regimens

We use  $\underline{g}_m$  to denote a general maximum weight gain dietary intervention regimen beginning at time  $m$ . Mathematically  $\underline{g}_m$  is a collection of functions  $\underline{g}_m = \{g_k[\bar{a}(k-1), \bar{l}(k)]; k = m, \dots, K\}$ . Under a regimen  $\underline{g}_m$ , a subject follows his own observed diet history prior to  $m$  and then, for  $K \geq k \geq m$ ,  $g_k[\bar{a}(k-1), \bar{l}(k)]$  is a non-negative function that specifies the increase in maximum BMI to be allowed at time  $k$  for a

subject with past BMI and covariate history  $[\bar{a}(k-1), \bar{l}(k)]$ . See the definition in the following paragraph for a precise statement. We use  $g$  as a shorthand for a regimen  $g_{\underline{g}_0}$  beginning at time 0. Note that any regimen  $g = g_{\underline{g}_0} = \{g_k[\bar{a}(k-1), \bar{l}(k)]; k=0, \dots, K\}$  is naturally associated with a particular regimen  $g_m$ : the regimen  $g_m = \{g_k[\bar{a}(k-1), \bar{l}(k)]; k=m, \dots, K\}$  where one follows his observed diet up till time  $m$  and then follows regimen  $g_m$  using functions  $g_k[\bar{a}(k-1), \bar{l}(k)]$  specified by  $g$  for  $k \geq m$ . Therefore, we can define the following counterfactuals.

Let  $Y_m^g$  be a subject's utility measured at the end of follow-up when the counterfactual intervention  $g_m$  is followed. Similarly, let  $\bar{B}M\bar{I}_m^g(k)$ ,  $\bar{L}_m^g(k)$ ,  $BMI_{m,\max}^g(k)$ ,  $\bar{A}_m^g(k)$  be a subject's BMI, covariate, maximum BMI and  $A$ -history through  $k$  under  $g_m$ . Note  $\bar{B}M\bar{I}_m^g(k) \in \bar{L}_m^g(k)$ . Then we have the following formal definition.

*Definition of a general time  $m$  maximum weight gain dietary intervention regimen  $g_m$* : The subject follows his observed diet up to time  $m$  and from month  $m$  onwards, the subject is weighed every day: (i) if  $A(m) = BMI(m+1) - BMI_{\max}(m) \geq g_m[\bar{A}(m-1), \bar{L}(m)]$ , the subject's caloric intake is restricted until the subject's BMI falls below  $BMI_{\max}(m) + g_m[\bar{A}(m-1), \bar{L}(m)]$ ; (ii) for  $m+1 \leq k \leq K$  if (a)  $A_m^g(k) \equiv BMI_m^g(k+1) - BMI_{m,\max}^g(k) \geq g_k[A_m^g(k-1), \bar{L}_m^g(k)]$ , the subject's caloric intake is restricted until the subject's BMI falls below  $BMI_{m,\max}^g(k) + g_k[A_m^g(k-1), \bar{L}_m^g(k)]$ ; (b) if his BMI is less than  $BMI_{m,\max}^g(k) + g_k[A_m^g(k-1), \bar{L}_m^g(k)]$ , the subject is allowed to eat as he pleases without any intervention.

Note, by definition,  $\bar{L}_m^g(k)$  equals  $\bar{L}_m(k)$  and  $A_m^g(k-1)$  equals  $A_m(k-1)$  for  $k \leq m$ . Furthermore, given a regimen  $g = g_{\underline{g}_0}$ , we say a subject's observed data are consistent with following the associated regimen  $g_m$  if and only if  $A_m^g(k) \leq g_k[A_m^g(k-1), \bar{L}_m^g(k)]$  for  $k \geq m$ . It follows that if a subject's observed data are consistent with following the associated regimen  $g_m$ , then subject's observed data are consistent with following the associated regimen  $g_k$  for any  $k > m$ .

If for all  $k \geq m$ ,  $g_k[\bar{a}(k-1), \bar{l}(k)]$  is a constant  $a(k)$  that does not depend on  $(\bar{a}(k-1), \bar{l}(k))$ , the regimen  $g_m$  is said to be non-dynamic or static and is written  $g_m = \underline{a}(m)$ . Otherwise it is dynamic. An intervention that allowed a BMI gain of 0.1/12 per month (that is, of one per decade) starting at time 0 (age 18 years) is the regimen  $g_{\underline{g}_0} = \underline{a}(0)$  with each  $a(m) = 0.1/12$ . A dynamic intervention starting at time 0 that allows a BMI gain of 0.1/12 per month in subjects free of hypertension, Db, hyperlipidemia or clinical CHD, but of only 0.05/12 per month once a subject developed one of these risk factors is a dynamic regimen  $g_{\underline{g}_0}$  with  $g_k[\bar{a}(k-1), \bar{l}(k)] = 0.1/12$  if  $\bar{l}(k)$  indicates a subject is free at  $k$  of hypertension, Db, hyperlipidemia or clinical CHD and  $g_k[\bar{a}(k-1), \bar{l}(k)] = 0.05/12$  otherwise.

The expected value  $E[Y_0^g]$  is our parameter of interest associated with the regimen  $g$ : the expected utility had we placed in 1950 all 18-year-old non-smoking American men on the maximum weight gain intervention regimen  $g$ .

Let  $\lfloor t \rfloor$  denote the smallest integer less than or equal to  $t$  and define  $b_+ = b$  if  $b \geq 0$  and  $b_+ = 0$  if  $b < 0$ . Note because

data are only obtained monthly, for any non-negative real number  $t$ ,  $A(t) = A(\lfloor t \rfloor)$  and  $L(t) = L(\lfloor t \rfloor)$ . Given a regimen  $g$ , let  $A_\Delta^g(t) = [A(\lfloor t \rfloor) - g_{\lfloor t \rfloor}[\bar{A}(\lfloor t \rfloor - 1), \bar{L}(\lfloor t \rfloor)]]_+ + [BMI(\lfloor t \rfloor + 1) - \{BMI_{\max}(\lfloor t \rfloor) + g_{\lfloor t \rfloor}[\bar{A}(\lfloor t \rfloor - 1), \bar{L}(\lfloor t \rfloor)]]_+$  so  $A_\Delta^g(t) = 0$  for all  $t$  if and only if a subject's observed data are consistent with following regimen  $g$  from time 0. When  $A_\Delta^g(t) \neq 0$ ,  $A_\Delta^g(t)$  measures how much greater one's observed weight gain is than the maximum prescribed by  $g$ . Define

$$X_m^g(\psi) = m + \int_m^x \exp\{\omega(A_\Delta^g(t), \bar{A}(t^-), \bar{L}(t), \psi)\} dt \quad \text{if } X > m \tag{49}$$

$$X_m^g(\psi) = X \quad \text{if } X \leq m \tag{50}$$

$$Y_j^g(\beta, \psi) = Y - \sum_{m=j}^K \gamma_m[A_\Delta^g(m), \bar{A}(m-1), \bar{L}(m), X_m(\psi), \beta] \tag{51}$$

where the functions  $\omega(a(t), \bar{a}(t^-), \bar{l}(t), \psi)$  and  $\gamma_m(a(m), \bar{a}(m-1), \bar{l}(m), \psi)$  are again known functions satisfying  $\omega(a(t), \bar{a}(t^-), \bar{l}(t), \psi) = 0$  if  $a(t) = 0$  or  $\psi = 0$  and  $\gamma_m(a(m), \bar{a}(m-1), \bar{l}(m), \beta) = 0$  if  $a(m) = 0$  or  $\beta = 0$ .

Given a regimen  $g$ , we say that (49) and (50) is a correctly specified SNFTM for  $X_m^g$  and (51) is a correctly specified SNMM for  $Y_m^g|X_m^g$  with true parameters  $(\beta^*, \psi^*)$  when there exists some  $(\beta^*, \psi^*)$  such that, for each  $m$ ,

*Assumption (i)*:  $X_m^g$  and  $X_m^g(\psi^*)$  have the same conditional distribution given  $(A_\Delta^g(j), \bar{A}(j-1), \bar{L}(j))$  and

*Assumption (ii)*:

$$E[Y_m^g|A_\Delta^g(m), \bar{A}(m-1), \bar{L}(m), X_m^g = x] = E[Y_m^g(\beta^*, \psi^*)|A_\Delta^g(m), \bar{A}(m-1), \bar{L}(m), X_m^g(\psi^*) = x] \tag{52}$$

Recall  $\bar{A}(m-1)$  is a function of  $\bar{L}(m)$  and thus its appearance in the conditioning event is redundant. Define

$$\Xi^g(m) = 1 \Leftrightarrow BMI(m+1) \geq BMI_{\max}(m) + g_m[\bar{A}(m-1), \bar{L}(m)] \tag{53}$$

so  $A_\Delta^g(m) > 0$  implies  $\Xi^g(m) = 1$ .

Given a regimen  $g$ , let the  $RC^g$  assumption be the  $RC$  assumption but with  $X_m^g$ ,  $Y_m^g$ ,  $\Xi^g(m)$  replacing their counterparts without  $g$  and  $A_\Delta^g$  replacing  $A$ . Let  $CD^g$  be the  $CD$  assumption but with  $X_m^g$  replacing  $X_m$  and 'time  $m$  dietary intervention' replaced by the ' $g_m$  dietary intervention.' Henceforth we assume the  $CD^g$  and the  $RC^g$  hold for all regimens  $g$ .

Suppose we carry out  $g$ -estimation as in the section Estimation of the effect of the 'maintain baseline weight intervention' except with  $X_m^g(\psi)$ ,  $Y_m^g(\beta, \psi)$ ,  $\Xi^g(m)$  replacing their counterparts without  $g$  and  $A_\Delta^g$  replacing  $A$ . Then results of Robins (4) imply that, under the  $RC^g$  and  $CD^g$  assumptions, if the model

$$E[A_\Delta^g|\bar{L}(m), \bar{A}(m-1), \Xi^g(m) = 1] = \alpha^T W(m)$$

is correct, and our SNFTM for  $X_m^g$  and SNMM for  $Y_m^g|X_m^g$  are correctly specified, then  $\tilde{\psi}$ ,  $\tilde{\beta}(\tilde{\psi})$  and  $n^{-1} \sum_i Y_{0i}^g[\tilde{\beta}(\tilde{\psi}), \tilde{\psi}]$

are CAN for  $\psi^*$ ,  $\beta^*$  and the parameter of interest  $E[Y_0^S]$ , respectively, provided  $(\beta^*, \psi^*)$  are identified and we choose  $Q_m(\beta)$  linear in  $Y_m(\beta)$ .

## Measurement error

In studies of the effect of a time-independent exposure, random exposure measurement error generally leads to bias toward the null and loss of power. However, the consequences of random exposure measurement error are much more complex in longitudinal studies of a time-dependent exposure in the presence of time-varying confounders. Specifically, in such a study, exposure history prior to time  $t$  needs to be considered as a potential confounder for the effect of exposure at  $t$ , even under the sharp null hypothesis of no causal effect of exposure at any time on the outcome  $Y$ . As random measurement error in a confounder can cause bias in any direction, random error in recorded BMI can, in principle, cause bias even under the null! See Robins<sup>13</sup>. Furthermore this random error should be seen as including not only errors in measurement of BMI but also short-term fluctuations in BMI due to illness, a New Years resolution to lose weight, etc. These random fluctuations in BMI may have little effect on eventual mortality, but they can easily obscure the actual trend in someone's BMI for periods of up to a year. Thus, if we use a monthly scale of analysis as described above, the random fluctuations in BMI may dominate any trend within a subject. Further given that past BMI must be controlled for in the regression models for current BMI used in g-estimation, the true correlation between past and present BMI trends within a person will be obscured by random fluctuations, which can even result in bias away from the null. This can occur when the confounding effect of past trends in BMI are inadequately controlled due to the random mismeasurement in past BMI. What to do?

One approach would be to specify a complex statistical model for the relationship between true and mismeasured BMI. At present I tend to seriously doubt the robustness of such an approach owing to inevitable model misspecification.

The alternative is to increase the 'time' between measurements used in the analysis from say 1 month up to as high as 5–6 years. By increasing the time between measurements, the problem of random fluctuations in BMI is markedly reduced, as the BMI signal (the true difference between measurement occasions) is made much greater, whereas the random fluctuations may not increase or may even decrease if the fluctuations are autocorrelated on a timescale of a few months. The drawback of increasing the 'time' between measurements in the analysis is that this can lead to poorer control of the confounding attributable to evolving time-varying factors. As an example, because the temporal ordering of events between the measurement times used in the analysis is lost, the confounding effect of changes in exercise may be incorrectly attributed to a causal effect of BMI.

At present I would recommend repeating one's analysis using a number of different between measurements 'times'

and report all results. In this way, the sensitivity of one's conclusions to the choice of the 'time' between measurements will be known. If important, this sensitivity will stimulate further discussion and the development of better analytic methods.

## Conflict of interest

The author declared no financial interests.

## References

- 1 Willett WC, Dietz WH, Colditz GA. Guidelines for healthy weight. *N Engl J Med* 1999; **341**: 427–434.
- 2 Robins JM. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 1992; **79**: 321–334.
- 3 Robins JM, Wasserman L. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In: Geiger D, Shenoy P (eds). *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, Providence Rhode Island, 1–3 August 1997. Morgan Kaufmann: San Francisco, 1997, pp 409–420.
- 4 Robins JM. Association, causation, and marginal structural models. *Synthese* 1999; **121**: 151–179.
- 5 Robins JM, Hernan MA, Siebert U. Effects of multiple interventions. In: Ezzati M, Lopez AD, Rodgers A, Murray CJL (eds). *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, vol 1. World Health Organization: Geneva, 2004, pp 2191–2230.
- 6 Hernán MA, Hernandez Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–625.
- 7 Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat* 1994; **23**: 2379–2412.
- 8 Robins JM. Optimal structural nested models for optimal sequential decisions. In: Lin DY, Heagerty P (eds). *Proceedings of the Second Seattle Symposium on Biostatistics*. Springer-Verlag: New York, 2004.
- 9 Robins JM. Causal inference from complex longitudinal data. In: Berkane M (ed). *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)*. Springer-Verlag: New York, 1997, pp 69–117.
- 10 Murphy SA. Optimal dynamic treatment regimes. *J R Stat Soc Ser B* 2003; **65**: 331–366.
- 11 Robins JM, Scharfstein D, Rotnitzky A. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D (eds). *Statistical Models in Epidemiology: the Environment and Clinical Trials*. Springer-Verlag: New York, 1999, pp 1–94.
- 12 Lok JJ, Gill RD, van der Vaart AW, Robins JM. Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Stat Neerl* 2001; **58**: 271–295.
- 13 Robins JM. General methodological considerations. *J Econom* 2003; **112**: 89–106.
- 14 Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Glymour C, Cooper G (eds). *Computation, Causation, and Discovery*. AAAI Press/The MIT Press: Menlo Park, CA; Cambridge, MA, 1999, pp 349–405.
- 15 Robins JM. Analytic methods for estimating HIV treatment and cofactor effects. In: Ostrow DG, Kessler R (eds). *Methodological Issues of AIDS Mental Health Research*. Plenum Publishing: New York, 1993, pp 213–290.
- 16 Joffe MM, Hoover DR, Jacobson LP, Kingsley L, Chmiel JS, Fischer BR *et al*. Estimating the effect of Zidovudine on Kaposi's sarcoma from observational data using a rank preserving failure time model. *Stat Med* 1998; **17**: 1073–1102.

Appendix 1

A formal definition of a joint SNFTM for  $X_m$  and a SNMM for  $Y_m|X_m$

The definition here is the alternative, more intuitive and more general definition mentioned in the main text. The equivalence with the definitions in the main text is proved below.

We first consider the uncensored case. The observed data are  $O = \bar{A}(K), \bar{L}(K + 1), X, Y$ , where  $X$  is a continuous time to event variable and  $Y$  is measured at  $K + 1$ . The counterfactual data are  $(X_m, Y_m), m = 0, \dots, K + 1$ , denoting  $X$  and  $Y$  under treatment regimens where one experiences his observed treatment  $\bar{A}(m - 1)$  up to  $m$  and then receives no treatment (treatment level 0) thereafter. We make the assumption that  $X_{K+1} = X, Y_{K+1} = Y$ . The covariate  $L(k)$  precedes  $A(k)$  which precedes  $L(k + 1)$ .

The function  $x_m^\dagger(x, \bar{L}(m), \bar{A}(m)) = S_{X_m|\bar{L}(m), \bar{A}(m)}^{-1}\{S_{X_{m+1}|\bar{L}(m), \bar{A}(m)}(x)\}$  is a counterfactual conditional quantile-quantile function, where  $S$  and  $S^{-1}$  denote a survivor function and its inverse. It is a standard result that  $x_m^\dagger(x, \bar{L}(m), \bar{A}(m))$  is the unique function for which  $X_m^* \equiv x_m^\dagger(X_{m+1}, \bar{L}(m), \bar{A}(m))$  and  $X_m$  have the same conditional distribution, that is,

$$X_m^*|\bar{L}(m), \bar{A}(m) \sim X_m|\bar{L}(m), \bar{A}(m) \tag{54}$$

Define  $X_{K+1}^\dagger = X$  and then recursively define  $X_m^\dagger \equiv x_m^\dagger(X_{m+1}^\dagger, \bar{L}(m), \bar{A}(m))$ . Robins and Wasserman<sup>3</sup> proved the following

Theorem A1:

$$X_m|\bar{L}(m), \bar{A}(m) \sim X_m^\dagger|\bar{L}(m), \bar{A}(m) \tag{55}$$

where we silently take such equations to hold for all  $m = 0, \dots, K$ .

Furthermore, Robins and co-workers<sup>11,14</sup> and Lok<sup>12</sup> proved the function  $x_m^\dagger$  is unique. That is if the above equation holds for with  $X_m^\dagger$  replaced by some  $H_m = h_m(H_{m+1}, \bar{L}(m), \bar{A}(m))$  and  $H_{K+1} = X$ , then the function  $h_m$  must be the function  $x_m^\dagger$ .

An SNFTM for  $X_m$  assumes

$$x_m(X_{m+1}, \bar{L}(m), \bar{A}(m); \psi^*) = x_m^\dagger(X_{m+1}, \bar{L}(m), \bar{A}(m)) \tag{56}$$

for a known function  $x_m(x, \bar{L}(m), \bar{A}(m); \psi)$  satisfying  $x_m(x, \bar{L}(m), \bar{A}(m), \psi) = x$  if  $\psi = 0$  or  $A(m) = 0$  with  $\psi^*$  an unknown parameter vector.

It follows immediately that

$$X_m(\psi^*)|\bar{L}(m), \bar{A}(m) \sim X_m^\dagger|\bar{L}(m), \bar{A}(m) \tag{57}$$

with

$$\begin{aligned} X_{K+1}(\psi^*) &= X \quad \text{and} \quad X_m(\psi^*) \\ &\equiv x_m(X_{m+1}(\psi^*), \bar{L}(m), \bar{A}(m); \psi^*) \end{aligned} \tag{58}$$

The uniqueness of  $X_m^\dagger$  implies that SNFTMs as defined in the text are also SNFTMs as defined here.

Recall  $X_m^* \equiv x_m^\dagger(X_{m+1}, \bar{L}(m), \bar{A}(m))$  and define

$$\begin{aligned} \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x) &\equiv E[Y_{m+1}|\bar{A}(m), \bar{L}(m), X_m^* = x] \\ &- E[Y_m|\bar{A}(m), \bar{L}(m), X_m = x] \end{aligned} \tag{59}$$

which is equivalent to

$$\begin{aligned} E[Y_{m+1} - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), X_m^*)|\bar{A}(m), \bar{L}(m), X_m^* = x] \\ = E[Y_m|\bar{A}(m), \bar{L}(m), X_m = x] \end{aligned} \tag{60}$$

Define  $Y_{K+1}^\dagger = Y$  and then recursively define  $Y_m^\dagger = Y_{m+1}^\dagger - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), X_m^\dagger)$ .

We prove the following theorem below.

Theorem A2:

$$E[Y_m^\dagger|\bar{L}(m), \bar{A}(m), X_m^\dagger = x] = E[Y_m|\bar{L}(m), \bar{A}(m), X_m = x] \tag{61}$$

Furthermore the function  $\gamma_m^\dagger$  is unique. That is if the above equation holds with  $Y_m^\dagger$  replaced by some  $H_m = H_{m+1} - h_m(\bar{A}(m), \bar{L}(m), X_m^\dagger)$  and  $H_{K+1} = Y$ , then the function  $h_m$  must be the function  $\gamma_m^\dagger$ .

An additive SNMM for  $Y_m|X_m$  assumes

$$\gamma_m(\bar{A}(m), \bar{L}(m), x; \beta^*) = \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x) \tag{62}$$

for a known function  $\gamma_m(\bar{A}(m), \bar{L}(m), x; \beta)$  satisfying  $\gamma_m(\bar{A}(m), \bar{L}(m), x; \beta) = 0$  if  $\beta = 0$  or  $A(m) = 0$  with  $\beta^*$  an unknown parameter vector.

It follows immediately that

$$\begin{aligned} E[Y_m(\beta^*, \psi^*)|\bar{L}(m), \bar{A}(m), X_m(\psi^*) = x] \\ = E[Y_m|\bar{L}(m), \bar{A}(m), X_m = x] \end{aligned} \tag{63}$$

with

$$\begin{aligned} Y_{K+1}(\beta^*, \psi^*) &= Y \quad \text{and} \quad Y_m(\beta^*, \psi^*) \equiv Y_{m+1}(\beta^*, \psi^*) \\ &- \gamma_m(\bar{A}(m), \bar{L}(m), X_m(\psi^*); \beta^*) \end{aligned} \tag{64}$$

The uniqueness of  $\gamma_m^\dagger$  implies that an additive SNMM for  $Y_m|X_m$  as defined in the text is equivalent to the additive SNMM for  $Y_m|X_m$  as defined here.

Proof of Theorem A2: By backward induction.

Case 1:  $m = K$ ;

$$\begin{aligned} E[Y_K^\dagger|\bar{L}(K), \bar{A}(K), X_K^\dagger = x] \\ = E[Y_{K+1} - \gamma_K^\dagger(\bar{A}(K), \bar{L}(K), X_K^\dagger)|\bar{A}(K), \bar{L}(K), X_K^\dagger = x] \\ = E[Y_{K+1} - \gamma_K^\dagger(\bar{A}(K), \bar{L}(K), X_K^*)|\bar{A}(K), \bar{L}(K), X_K^* = x] \\ = E[Y_K|\bar{A}(K), \bar{L}(K), X_K = x] \end{aligned}$$

where the first equality uses the definition of  $Y_K^\dagger$  and that  $Y_{K+1} = Y = Y_{K+1}^\dagger$ , the second uses that  $X_K^* = X_K^\dagger$  by  $X_{K+1} = X = X_{K+1}^\dagger$ , and the third is the definition of  $\gamma_K^\dagger(\bar{A}(K), \bar{L}(K), X_K^*)$ .

Case 2: Assume true for  $m$ . We prove true for  $m + 1$ .

$$\begin{aligned} E[Y_m^\dagger|\bar{L}(m), \bar{A}(m), X_m^\dagger = x] \\ = E[Y_{m+1}^\dagger - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), X_m^\dagger)|\bar{L}(m), \bar{A}(m), X_m^\dagger = x] \text{ (by def of } Y_m^\dagger) \\ = E\left\{ \int E[Y_{m+1}^\dagger|\bar{L}(m+1), \bar{A}(m+1), X_{m+1}^\dagger = u] dF_{X_{m+1}^\dagger}(u|\bar{L}(m+1), \bar{A}(m+1), X_m^\dagger = x) \right\} \\ [\bar{L}(m), \bar{A}(m), X_m^\dagger = x] \\ - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x) \end{aligned}$$

(by the laws of probability)

$$= E \left[ \left\{ \int E[Y_{m+1} | \bar{L}(m+1), \bar{A}(m+1), X_{m+1} = u] dF_{X_{m+1}^\dagger} (u | \bar{L}(m+1), \bar{A}(m+1), X_m^* = x) \right\} \right. \\ \left. - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x) \right]$$

(by the induction assumption)

$$= E \left[ \left\{ \int E[Y_{m+1} | \bar{L}(m+1), \bar{A}(m+1), X_{m+1} = u] \right. \right. \\ \left. \left. dF_{X_{m+1}^\dagger} (u | \bar{L}(m+1), \bar{A}(m+1), X_m^*(X_{m+1}, \bar{L}(m), \bar{A}(m)) = x) \right\} \right] \\ - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x)$$

(by  $\{X_{m+1}, \bar{L}(m+1), \bar{A}(m+1)\}$  and  $\{X_{m+1}^\dagger, \bar{L}(m+1), \bar{A}(m+1)\}$  having identical distributions)

$$= E[Y_{m+1} | \bar{L}(m), \bar{A}(m), X_m^*(X_{m+1}, \bar{L}(m), \bar{A}(m)) = x] - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x)$$

(by the laws of probability)

$$= E[Y_{m+1} | \bar{L}(m), \bar{A}(m), X_m^* = x] - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x)$$

(by the definition of  $X_m^*$ )

$$= E[Y_m | \bar{A}(m), \bar{L}(m), X_m = x]$$

(by the definition of  $\gamma_m^\dagger(\bar{A}(m), \bar{L}(m), x)$ ).

Uniqueness is proved as in Robins<sup>8</sup> and Lok *et al.*<sup>12</sup> and is omitted.

An additive SNMM for  $Y_m | X_m$  may not be appropriate for analyzing censored data due to administrative censoring of  $X$  at time  $K$  as discussed in the text. As indicated in the section Censoring, our approach requires that we consider a broader class of SNMM models, which we now describe.

Consider a collection of functions  $c_m^\dagger(x, \bar{A}(m), \bar{L}(m))$  indexed by  $m$  and define  $C_m^* = c_m^\dagger(X_m^*, \bar{A}(m), \bar{L}(m))$ ,  $C_m^\dagger = c_m^\dagger(X_m^\dagger, \bar{A}(m), \bar{L}(m))$ , and  $C_m = c_m^\dagger(X_m, \bar{A}(m), \bar{L}(m))$ . For fixed  $\bar{A}(m), \bar{L}(m)$ ,  $c_m^\dagger(x, \bar{A}(m), \bar{L}(m))$  need not be a 1-1 function of  $x$ .

Redefine

$$\gamma_m^\dagger(\bar{A}(m), \bar{L}(m), c) \equiv E[Y_{m+1} | \bar{A}(m), \bar{L}(m), C_m^* = c] \\ - E[Y_m | \bar{A}(m), \bar{L}(m), C_m = c] \quad (65)$$

which is equivalent to

$$E[Y_{m+1} - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), C_m^*) | \bar{A}(m), \bar{L}(m), C_m^* = c] \\ = E(Y_m | \bar{A}(m), \bar{L}(m), C_m = c) \quad (66)$$

Define  $Y_{K+1}^\dagger = Y$  and then recursively redefine  $Y_m^\dagger = Y_{m+1}^\dagger - \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), C_m^\dagger)$ . We have the following theorem.

*Theorem A3:*

$$E[Y_m^\dagger | \bar{L}(m), \bar{A}(m), C_m^\dagger = c] = E[Y_m | \bar{L}(m), \bar{A}(m), C_m = c] \quad (67)$$

Furthermore the function  $\gamma_m^\dagger$  is unique. That is, if the above equation holds with  $Y_m^\dagger$  replaced by some  $H_m = H_{m+1} - h_m(\bar{A}(m), \bar{L}(m), C_m^\dagger)$  and  $H_{K+1} = Y$ , then the function  $h_m$  must be the function  $\gamma_m^\dagger$ .

*Proof of A.3:* We only describe where the proof differs from that of its special case Theorem A.2. The proof is essentially identical except for the replacement of  $X_m$  by  $C_m$ ,  $x$  by  $c$  and  $x_m^\dagger(X_{m+1}, \bar{L}(m), \bar{A}(m))$  by

$$c_m^\dagger(x_m^\dagger(X_{m+1}, \bar{L}(m), \bar{A}(m)), \bar{L}(m), \bar{A}(m)) \quad (68)$$

An additive SNMM for  $Y_m | C_m$  assumes

$$\gamma_m(\bar{A}(m), \bar{L}(m), c; \beta^*) = \gamma_m^\dagger(\bar{A}(m), \bar{L}(m), c) \quad (69)$$

for a known function  $\gamma_m(\bar{A}(m), \bar{L}(m), c; \beta)$  satisfying  $\gamma_m(\bar{A}(m), \bar{L}(m), c; \beta) = 0$  if  $\beta = 0$  or  $A(m) = 0$  with  $\beta^*$  an unknown parameter vector.

Then given a SNFTM  $x_m(X_{m+1}, \bar{L}(m), \bar{A}(m); \psi^*)$  for  $X_m$  and a function  $c_m^\dagger(x, \bar{A}(m), \bar{L}(m))$  that may depend on the functions  $x_j^\dagger(\cdot, \cdot, \cdot)$ ,  $j \geq m$ , suppose we can define a parametrized class of functions  $c_m(x, \bar{A}(m), \bar{L}(m), \psi)$  satisfying

$$c_m(x, \bar{A}(m), \bar{L}(m), \psi^*) = c_m^\dagger(x, \bar{A}(m), \bar{L}(m)) \quad (70)$$

For example, in the section on censoring in the main text, we took  $c_m(x, \bar{A}(m), \bar{L}(m), \psi) = \min(x, K_m(\psi))$ .

Then defining  $C_m(\psi) = c_m(X_m(\psi), \bar{A}(m), \bar{L}(m), \psi)$ , we have

$$E[Y_m(\beta^*, \psi^*) | \bar{L}(m), \bar{A}(m), C_m(\psi^*) = x] \\ = E[Y_m | \bar{L}(m), \bar{A}(m), C_m = x] \quad (71)$$

with

$$Y_{K+1}(\beta^*, \psi^*) = Y \text{ and } Y_m(\beta^*, \psi^*) \equiv Y_{m+1}(\beta^*, \psi^*) \\ - \gamma_m(\bar{A}(m), \bar{L}(m), C_m(\psi^*); \beta^*) \quad (72)$$

## Appendix 2

*Estimation of effects with the parametric g-formula and IPTW when a sufficiently long MLP exists*

In this section, we show that the parametric g-formula and IPTW can be used to estimate certain causal effects when there exists a sufficiently long MLP. We begin with a preliminary discussion of these two methods of estimation.

*Preliminaries.* In this preliminary discussion, we assume that, as in the section A locally rank-preserving SNM, there is neither confounding by preclinical disease nor an MLP. Specifically we assume, for each regimen  $g$ , the CO<sup>g</sup> assumption that, for each  $j$ ,  $(Y_0^g, X_0^g) \perp\!\!\!\perp A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{O}(j-1), \Xi^g(j) = 1$  holds, with  $\Xi^g(m)$  defined in Equation (53).

*Recoding:* Without loss of generality, we henceforth redefine (that is, recode)  $\bar{L}(j)$  such that  $\Xi^g(j)$  is now one of the components of  $\bar{L}(j)$  but we remove from  $\bar{L}(j)$  the components corresponding to  $X$ , that is, the components  $(XI(X \leq j), I(X \leq j))$ . Then we can write the  $CO^g$  assumption as

$$CO^g : (Y_j^g, X_j^g) \prod A_{\Delta}^g(j) | \bar{L}(j), \bar{A}_{\Delta}^g(j-1) = \bar{0}(j-1), (XI(X \leq j), I(X \leq j)) \quad (73)$$

as, from these definitions,  $\Xi^g(j) = 0$  implies  $A_{\Delta}^g(j) = 0$ . The  $CO^g$  assumption implies

$$(Y_0^g, X_0^g) \prod A_{\Delta}^g(j) | \bar{L}(j), \bar{A}_{\Delta}^g(j-1) = \bar{0}(j-1), (XI(X \leq j), I(X \leq j)) \quad (74)$$

as  $\bar{A}_{\Delta}^g(j-1) = \bar{0}(j-1)$  implies  $(Y_j^g, X_j^g) = (Y_0^g, X_0^g)$ . This last equation is the standard definition of no unmeasured confounding given  $(\bar{L}(j), (XI(X \leq j), I(X \leq j)))$  for the effect of  $A_{\Delta}^g(j)$  on the counterfactuals  $Y_0^g, X_0^g$ . Let  $\lambda(u|\cdot) = \lim_{h \rightarrow 0} \text{pr}[u \leq X < u+h | \cdot, u \leq X] / h$  be the conditional hazard of  $X$  given the information in  $\bullet$ .

Robins<sup>4,15</sup> proves that Equation (74) implies that  $S_{X_0^g}(u) \equiv \text{pr}(X_0^g > u)$  is identified through

$$S_{X_0^g}(u) = \int \dots \int \exp \left\{ - \int_0^u \lambda(t | \bar{L}(t), \bar{A}_{\Delta}^g(t) = \bar{0}) dt \right\} \quad (75)$$

$$\times \prod_{m=0}^{m=\lfloor u \rfloor} dF[L(m) | \bar{L}(m-1), \bar{A}_{\Delta}^g(m-1) = \bar{0}, X > m] = E[I\{X > u\} I\{\bar{A}_{\Delta}^g(u) = \bar{0}\} \mathbb{W}^{g,*}(u)] \quad (76)$$

with

$$\mathbb{W}^{g,*}(u) = 1 / \prod_{m=0}^{m=\lfloor u \rfloor} \text{pr}[A_{\Delta}^g(m) = 0 | \bar{L}(m), \bar{A}_{\Delta}^g(m-1), X > m] \quad (77)$$

where the first formula for  $S_{X_0^g}(u)$  is referred to as the g-computation algorithm formula (g-formula, for short) and the second formula as the IPTW formula. To shorten the formulae, we have written  $\bar{0}$  as a shorthand for  $\bar{0}(t)$  when the time  $t$  is clear. In fact Robins<sup>4,15</sup> shows that the assumption

$$(Y_0^g, X_0^g) \prod A_{\Delta}^g(j) | \bar{L}(j), \bar{A}_{\Delta}^g(j-1) = \bar{0}(j-1), X > j \quad (78)$$

which is implied by the assumption of Equation (74), suffices to establish the identifying formulae. To estimate  $S_{X_0^g}(u)$ , we can use either the parametric g-formula estimator that replaces the unknowns  $\lambda(t | \bar{L}(t), \bar{A}_{\Delta}^g(t) = \bar{0})$  and  $f[L(m) | \bar{L}(m-1), \bar{A}_{\Delta}^g(m-1) = \bar{0}, X > m]$  in the first formula by estimates based on parametric models or the IPTW estimator that replaces the unknown  $\text{pr}[A_{\Delta}^g(m) = 0 | \bar{L}(m-1), \bar{A}_{\Delta}^g(m-1) = \bar{0}, X > m]$  in the second formula with a parametric estimate and the unknown expectation with a sample average. Both approaches are alternatives to g-estimation of structural nested models (SNMs).

Robins<sup>4,15</sup> proves  $E[Y_0^g]$  is identified under the assumption of Equation (74) by

$$E[Y_0^g] = \int_0^{K+1} dx \int \dots \int \lambda_X(x | \bar{L}(x), \bar{A}_{\Delta}^g(x) = \bar{0}) \times \exp \left\{ - \int_0^x \lambda_X(t | \bar{L}(t), \bar{A}_{\Delta}^g(t) = \bar{0}) dt \right\} \times \prod_{m=0}^{m=\lfloor x \rfloor} dF[L(m) | \bar{L}(m-1), \bar{A}_{\Delta}^g(m-1) = \bar{0}, X > m] \times \prod_{m=\lfloor x+1 \rfloor}^{K+1} dF[L(m) | \bar{L}(m-1), \bar{A}_{\Delta}^g(m-1) = \bar{0}, X = x] \times E[Y | \bar{L}(K+1), \bar{A}_{\Delta}^g(K) = \bar{0}, X = x] = E[Y I\{\bar{A}_{\Delta}^g(K) = \bar{0}\} \mathbb{W}^{g,*}]$$

$$\mathbb{W}^{g,*} = \mathbb{W}^{g,*}(X)$$

$$\times \left\{ 1 / \prod_{m=\lfloor X+1 \rfloor}^K \text{pr}[A_{\Delta}^g(m) = 0 | \bar{L}(m), \bar{A}_{\Delta}^g(m-1) = \bar{0}, X, X < m] \right\}$$

In the above formulae, we have assumed for simplicity that  $X$  has support on  $(0, K+1)$  so censoring for  $X$  is absent.

We next consider whether  $S_{X_0^g}(u)$  and  $E[Y_0^g]$  remain identified in the presence of confounding by preclinical disease and a sufficiently long MLP.

*Identification and estimation of  $S_{X_0^g}(u)$  :*

The following theorem establishes the identification of  $S_{X_0^g}(u)$ . First note under our recoding, the  $RC^g$  assumption becomes

$$RC^g : (Y_j^g, X_j^g) \prod A_{\Delta}^g(j) | \bar{L}(j), \bar{A}_{\Delta}^g(j-1), \bar{U}(j) = 0, (XI(X \leq j), I(X \leq j)) \quad (79)$$

*Theorem A4:* Given a regimen  $g$ , let a g-specific MLP satisfy the definition of an MLP of the section Estimation under a rank-preserving SNM for  $Y_m | X_m$  with  $X_m$  known, except with  $X_k$  and  $X_m$  replaced by  $X_k^g$  and  $X_m^g$  and  $A(m)$  replaced by  $A_{\Delta}^g(m)$ . Suppose  $A_{\Delta}^g(m)$  has a g-specific MLP of  $\chi$  months for its effect on  $X$  where  $\chi$  exceeds the time  $\varsigma$  in the  $CD^g$  assumption. Then, under the  $CD^g$  and  $RC^g$  assumptions,  $S_{X_0^g}(u)$  remains identified by both the g-formula and the IPTW formula when the recoded  $L(t)$  and  $A_{\Delta}^g(t)$  are redefined as  $L^\dagger(t)$  and  $A_{\Delta}^g(t)$  where

$$L^\dagger(t) = L(t - \chi), A_{\Delta}^g(t) = A_{\Delta}^g(t - \chi) \quad (80)$$

The theorem thus states that the identifying formulae are the usual g-formula and IPTW formula except we replace both the treatment variable  $A_{\Delta}^g(t)$  and the covariate variable  $L^\dagger(t)$  by their values  $\chi$  time units earlier. (For the IPTW formula, the transformation is applied to  $\mathbb{W}^{g,*}(u)$ .) It is important to emphasize that a similar transformation is not applied to  $X$ . Thus, the conditioning event

$\bar{L}(m-1), \bar{A}_\Delta^g(m-1) = \bar{0}, X > m$  transforms to  $\bar{L}(m-1-\chi), \bar{A}_\Delta^g(m-1-\chi) = \bar{0}, X > m$ .

*Proof of theorem:* It suffices to show Equation (78) holds when  $L(t)$  and  $A_\Delta^g(t)$  are replaced by  $L^\dagger(t)$  and  $A_\Delta^{g,\dagger}(t)$ . By RC<sup>g</sup>,  $(Y_j^g, X_j^g) \amalg A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{0}, \bar{U}(j) = \bar{0}, X > j$ . Thus,  $(Y_j^g, X_j^g) \amalg A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{0}, \bar{U}(j) = \bar{0}, X > j, X_j^g > j + \chi$ . By CD<sup>g</sup> and  $\chi > \varsigma$ ,  $(Y_j^g, X_j^g) \amalg A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{0}, X_j^g > j + \chi, X > j$ .

Thus  $(Y_{m-\chi}^g, X_{m-\chi}^g) \amalg A_\Delta^g(m-\chi) | \bar{L}(m-\chi), \bar{A}_\Delta^g(m-\chi-1) = \bar{0}, X > (m-\chi), X_{m-\chi}^g > m$  with  $m \equiv \chi + j$ .

Now the event  $X > (m-\chi)$  is the event  $X_{m-\chi}^g > (m-\chi)$ . Further, by  $\chi$  a g-specific MLP we also have the event  $X_{m-\chi}^g > m$  is the event  $X > m$ . Thus, we have  $(Y_{m-\chi}^g, X_{m-\chi}^g) \amalg A_\Delta^g(m-\chi) | \bar{L}(m-\chi), \bar{A}_\Delta^g(m-\chi-1) = \bar{0}, X > m$ . As, given  $\bar{A}_\Delta^g(m-\chi-1) = \bar{0}(m-\chi-1)$ , we have  $(Y_{m-\chi}^g, X_{m-\chi}^g) = (Y_0^g, X_0^g)$ , we conclude  $(Y_0^g, X_0^g) \amalg A_\Delta^g(m-\chi) | \bar{L}(m-\chi), \bar{A}_\Delta^g(m-\chi-1) = \bar{0}, X > m$ , which is exactly Equation (78) with  $L(t)$  and  $A_\Delta^g(t)$  replaced by  $L^\dagger(t)$  and  $A_\Delta^{g,\dagger}(t)$ , proving the theorem.

In contrast, under the conditions of the previous theorem,  $E[Y_0^g]$  is not identified because Equation (74), in contrast to Equation (78), fails to hold when  $L(t)$  and  $A_\Delta^g(t)$  are replaced by  $L^\dagger(t)$  and  $A_\Delta^{g,\dagger}(t)$ . Specifically, Equation (74) can be written as the conjunction of Equation (78),

$$(Y_0^g, X_0^g) \amalg A_\Delta^g(m) | \bar{L}(m), \bar{A}_\Delta^g(m-1) = \bar{0}, X, m > X > m - \chi + \varsigma \quad (81)$$

and

$$(Y_0^g, X_0^g) \amalg A_\Delta^g(m) | \bar{L}(m), \bar{A}_\Delta^g(m-1) = \bar{0}, X, m - \chi + \varsigma > X \quad (82)$$

We show below that under the conditions of the previous theorem, Equation (81) holds but Equation (82) does not when  $L(t)$  and  $A_\Delta^g(t)$  are replaced by  $L^\dagger(t)$  and  $A_\Delta^{g,\dagger}(t)$ . To show (81) we modify slightly the proof of Equation (78) as follows:

$$(Y_j^g, X_j^g) \amalg A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{0}, \bar{U}(j) = 0, X > j \text{ (by RC}^g\text{)}$$

$$\Rightarrow (Y_j^g, X_j^g) \amalg A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1)$$

$$= \bar{0}, \bar{U}(j) = \bar{0}, X > j, X_j^g, j + \varsigma < X_j^g < j + \chi$$

$$\Rightarrow (Y_j^g, X_j^g) \amalg A_\Delta^g(j) | \bar{L}(j), \bar{A}_\Delta^g(j-1)$$

$$= \bar{0}, X > j, X_j^g, j + \varsigma < X_j^g < j + \chi \text{ (by CD}^g\text{)}$$

$$\Rightarrow (Y_{m-\chi}^g, X_{m-\chi}^g) \amalg A_\Delta^g(m-\chi) | \bar{L}(m-\chi), \bar{A}_\Delta^g(m-\chi-1)$$

$$= \bar{0}, X > (m-\chi), X_{m-\chi}^g, m > X_{m-\chi}^g > m - \chi + \varsigma$$

$$\Rightarrow (Y_0^g, X_0^g) \amalg A_\Delta^g(m-\chi) | \bar{L}(m-\chi), \bar{A}_\Delta^g(m-\chi-1)$$

$$= \bar{0}, X_{m-\chi}^g > (m-\chi), X_{m-\chi}^g, m > X_{m-\chi}^g > m - \chi + \varsigma$$

$$\Rightarrow (Y_0^g, X_0^g) \amalg A_\Delta^g(m-\chi) | \bar{L}(m-\chi), \bar{A}_\Delta^g(m-\chi-1)$$

$$= \bar{0}, X, m > X > m - \chi + \varsigma$$

by the g-specific MLP assumption.

The proof of (82) fails because the event  $\bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{0}, \bar{U}(j) = \bar{0}, X > j, X_j^g, X_j^g < j + \varsigma$  is not the same event as  $\bar{L}(j), \bar{A}_\Delta^g(j-1) = \bar{0}, \bar{U}(j) = \bar{0}, X > j, X_j^g, X_j^g < j + \varsigma$ , under CD<sup>g</sup> because  $X_j^g < j + \varsigma$  does not imply  $\bar{U}(j) = \bar{0}$ .

*Proof that  $E[Y_0^g]$  is non-parametrically identified when a sufficiently long MLP exists.* In the section Intractable confounding in subgroups, we stated that  $E[Y_0^g]$  is non-parametrically identified under the conditions of the previous theorem with the regimen  $g$  in the theorem being the regimen that always assigns exposure zero. A proof follows.

Let  $IN, A^T, \Xi^T, Y_m^T, X_m^T$  be as defined in the section Intractable confounding in subgroups where we recall that because of the existence of the MLP of length  $\chi > \varsigma$ , all subjects with  $\varsigma < X_m < m + \varsigma$  have  $IN(m) = 1$ . First in Equations (78), (81) and (82) we replace  $(Y_0^g, X_0^g)$  by  $(Y_0^T, X_0^T)$ ,  $A_\Delta^g(m)$  by  $A^T(m-\chi)$ , and redefine  $L(m)$  as  $L(m-\chi)$  with the component  $\Xi(m)$  of  $L(m)$  being replaced by  $\Xi^T(m-\chi)$ . Equation (82) now holds trivially because with probability one  $m-\chi + \varsigma > X$  implies  $IN(m-\chi) = 1$  and thus  $\Xi^T(m-\chi) = 0$  and  $A^T(m-\chi) = 0$ . Furthermore, the proofs of Equations (78) and (81) go through as above with only minor notational changes. We therefore conclude that Equation (74) holds and thus that  $E[Y_0^g]$  is non-parametrically identified. The identifying IPTW formula is explicitly given by

$$E[Y_0^g] = E[YI\{A^T(K-\chi) = \bar{0}\} \mathbb{W}^{g,*}],$$

$$\begin{aligned} \{\mathbb{W}^{g,*}\}^{-1} &= \prod_{m=0}^{m=[X]} \text{pr}[A^T(m-\chi) \\ &= 0 | \bar{L}(m-\chi), \bar{A}^T(m-\chi-1) = 0, X > m] \\ &\times \left\{ \prod_{m=[X+1]}^K \text{pr}[A^T(m-\chi) = 0 | \bar{L}(m-\chi), \bar{A}^T(m-\chi-1) = 0, X] \right\} \end{aligned}$$

### Appendix 3

#### Optimal regimen models

Suppose we now wish to estimate the regimen  $g_{\text{opt}}$  that maximizes  $E[Y_0^g]$  over all regimens  $g$ . We will do so by specifying an optimal regimen SNMM and associated SNFTM.

To begin consider the dietary intervention  $a(k), g_{\text{opt},k+1}$  in which one follows his observed diet up to month  $k$ , allows a BMI increase of  $a(k)$  over his maximum previous BMI in month  $k$ , and follows the unknown optimal regimen  $g_{\text{opt}}$  thereafter. Let  $Y^{a(k),g_{\text{opt},k+1}}, X^{a(k),g_{\text{opt},k+1}}$  be the associated counterfactuals. When  $A(k) = a(k)$ , write  $g_{\text{opt},k+1}$  for the regimen  $A(k), g_{\text{opt},k+1}$ . Note  $X^{\xi_{\text{opt},K+1}} = X$ .

We will make the following assumptions:

*Optimal regimen RC assumption:*  $A(m)$  is statistically independent of  $(Y^{a(m),g_{\text{opt},m+1}}, X^{a(m),g_{\text{opt},m+1}})$  given  $\Xi(m) = 1, \bar{L}(m), \bar{A}(m-1)$  and  $\bar{U}(m) = \bar{0}(m) \geq 0$ .

*Optimal regimen CD assumption:*

$$X^{\xi_{\text{opt},m}} > m + \zeta \Rightarrow \bar{U}(m) = \bar{0}(m) \quad (83)$$

We next recursively define random variables  $X^{a(m), g_{opt,m+1}}(\psi)$  by the relationship that  $X^{g_{opt,K+1}}(\psi) = X$  and, for  $m = K, \dots, 0$ ,

$$\begin{aligned}
 X^{0(m), g_{opt,m+1}}(\psi) &= m + \exp\{\omega(a(m), \bar{A}(m-1), \bar{L}(m), \psi)\} \\
 &\quad \times (X^{a(m), g_{opt,m+1}}(\psi) - m) \\
 \text{if } 0 < X^{a(m), g_{opt,m+1}}(\psi) - m < 1 \\
 X^{0(m), g_{opt,m+1}}(\psi) &= X^{a(m), g_{opt,m+1}}(\psi) \\
 &\quad + \{\exp\{\omega(a(m), \bar{A}(m-1), \bar{L}(m), \psi)\} - 1\} \\
 \text{if } 1 < X^{a(m), g_{opt,m+1}}(\psi) - m \\
 X^{0(m), g_{opt,m+1}}(\psi) &= X^{a(m), g_{opt,m+1}}(\psi) \\
 \text{if } X^{a(m), g_{opt,m+1}}(\psi) < m
 \end{aligned}$$

These equations recursively define  $X^{a(m), g_{opt,m+1}}(\psi)$  in terms of the observed data, the regimen  $g_{opt,m+1}$  and the parameter vector  $\psi$  as can be verified by noting that these equations imply the following relationship between  $X^{a(m), g_{opt,m+1}}(\psi)$  and  $X^{g_{opt,m+1}}(\psi)$ .

$$\begin{aligned}
 X^{a(m), g_{opt,m+1}}(\psi) &= m \\
 &+ \frac{\exp\{\omega(A(m), \bar{A}(m-1), \bar{L}(m), \psi)\}}{\exp\{\omega(a(m), \bar{A}(m-1), \bar{L}(m), \psi)\}} (X^{g_{opt,m+1}}(\psi) - m)
 \end{aligned}$$

$$\text{if } 0 < X^{g_{opt,m+1}}(\psi) - m < 1, 0 < X^{a(m), g_{opt,m+1}}(\psi) < 1$$

$$\begin{aligned}
 X^{a(m), g_{opt,m+1}}(\psi) \\
 = X^{g_{opt,m+1}}(\psi) + \exp\{\omega(A(m), \bar{A}(m-1), \bar{L}(m), \psi)\} \\
 - \exp\{\omega(a(m), \bar{A}(m-1), \bar{L}(m), \psi)\}
 \end{aligned}$$

$$\text{if } 1 < X^{a(m), g_{opt,m+1}}(\psi) - m, 1 < X^{g_{opt,m+1}}(\psi) - m$$

$$\begin{aligned}
 X^{a(m), g_{opt,m+1}}(\psi) &= m + \exp\{\omega(A(m), \bar{A}(m-1), \bar{L}(m), \psi)\} \\
 (X_m^{g_{opt,m+1}}(\psi) - m)
 \end{aligned}$$

$$+ 1 - \exp\{\omega(a(m), \bar{A}(m-1), \bar{L}(m), \psi)\}$$

$$\text{if } 0 < X_m^{g_{opt,m+1}}(\psi) - m < 1, 1 < X^{a(m), g_{opt,m+1}}(\psi) - m$$

$$\begin{aligned}
 X^{a(m), g_{opt,m+1}}(\psi) &= m \\
 &+ \frac{\{\exp\{\omega(A(m), \bar{A}(m-1), \bar{L}(m), \psi)\} - 1\} + (X^{g_{opt,m+1}}(\psi) - m)\}}{\exp\{\omega(a(m), \bar{A}(m-1), \bar{L}(m), \psi)\}}
 \end{aligned}$$

$$\text{if } 0 < X^{a(m), g_{opt,m+1}}(\psi) - m < 1, 1 < X^{g_{opt,m+1}}(\psi) - m$$

We assume an optimal regimen SNFTM given by

$$X^{a(m), g_{opt,m+1}}(\psi^*) = X^{a(m), g_{opt,m+1}} \text{ wp1} \tag{84}$$

for an unknown value  $\psi^*$  of the vector  $\psi$ .

We also assume an optimal regimen SNMM

$$\begin{aligned}
 \gamma_m[a(m), \bar{a}(m-1), \bar{l}(m), x, \beta^*) \\
 \equiv E[Y^{a(m), g_{opt,m+1}} | \bar{L}_m = \bar{l}_m, \bar{A}_m = \bar{a}_m, X^{0(m), g_{opt,m+1}}(\psi^*) = x] \\
 - E[Y^{0(m), g_{opt,m+1}} | \bar{L}_m = \bar{l}_m, \bar{A}_m = \bar{a}_m, X^{0(m), g_{opt,m+1}} = x]
 \end{aligned} \tag{85}$$

Above  $\omega(a(t), \bar{a}(t-1), \bar{l}(t), \psi)$  and  $\gamma_m[a(m), \bar{a}(m-1), \bar{l}(m), x, \beta]$  are known functions satisfying  $\omega(a(t), \bar{a}(t-1), \bar{l}(t), \psi) = 0$  if  $a(t) = 0$  or  $\psi = 0$  and  $\gamma_m[a(m), \bar{a}(m-1), \bar{l}(m), \beta) = 0$  if  $a(m) = 0$  or  $\beta = 0$ .

The optimal regimen itself remains unknown. However, we show below that the following algorithm evaluated at the true  $(\beta^*, \psi^*)$  would find the optimal regimen  $g_{opt}$  under the following additional condition, that we henceforth assume to hold.

*Additional condition:* For each  $\bar{a}(m-1), \bar{l}(m), x, \beta, m$  the function  $\gamma_m^{opt}[a(m), \bar{a}(m-1), \bar{l}(m), x, \beta]$  is either everywhere zero or is strictly concave downward in  $a(m)$  on the support of  $A(m)$ .

*Optimal regimen algorithm:* Given any  $(\beta, \psi)$ , calculate  $g_{opt(\beta, \psi)} = \{g_{opt(\beta, \psi), m}[\bar{a}(m), \bar{l}(m)]; m = K, \dots, 0\}$  as follows.

Calculate  $X^{0(K), g_{opt,K+1}}(\psi)$ . Define

$$\begin{aligned}
 g_{opt(\beta, \psi), K}^*[\bar{A}; \bar{L}(K)] \\
 = I(X \leq K) \arg \max_{a(K)} [\gamma_K\{a(K), \bar{A}(K-1), \bar{L}(K), X, \beta\}] + I(X > K) \\
 \times \arg \max_{a(K)} E[\gamma_K\{a(K), \bar{A}(K-1), \bar{L}(K), X^{0(K), g_{opt,K+1}}(\psi), \beta\} \\
 | \bar{A}(K-1), \bar{L}(K), X > k]
 \end{aligned}$$

Calculate

$$g_{opt(\beta, \psi), K}[\bar{A}(K), \bar{L}(K)] = \min\{A(K), g_{opt(\beta, \psi), K}^*[\bar{A}(K-1), \bar{L}(K)]\}$$

Calculate  $X^{g_{opt(\beta, \psi), K}}(\psi) = X^{g_{opt(\beta, \psi), K}[\bar{A}(K), \bar{L}(K)]}$ .

Recursively for  $m = K-1, \dots, 0$ , calculate  $X^{0(m), g_{opt(\beta, \psi), m+1}}(\psi)$ ,

$$\begin{aligned}
 g_{opt(\beta, \psi), m}^*[\bar{A}(m-1), \bar{L}(m)] \\
 = I(X \leq m) \arg \max_{a(m)} E[\gamma_m\{a(m), \bar{A}(m-1), \bar{L}(m), X, \beta\}] + I(X > m) \\
 \times \arg \max_{a(m)} E[\gamma_m\{a(m), \bar{A}(m-1), \bar{L}(m), X^{0(m), g_{opt(\beta, \psi), m+1}}(\psi), \beta\} \\
 | \bar{A}(m-1), \bar{L}(m), X > m]
 \end{aligned}$$

Calculate

$$g_{opt(\beta, \psi), m}[\bar{A}(m), \bar{L}(m)] = \min\{A(m), g_{opt(\beta, \psi), m}^*[\bar{A}(m-1), \bar{L}(m)]\}.$$

Calculate

$$X^{g_{opt(\beta, \psi), m}}(\psi) = X^{g_{opt(\beta, \psi), m}[\bar{A}(m), \bar{L}(m)]}$$

Note that to carry out this algorithm we will need to be able to estimate

$$\begin{aligned}
 E[\gamma_m\{a(m), \bar{A}(m-1), \bar{L}(m), X^{0(m), g_{opt(\beta, \psi), m+1}}(\psi), \beta\} \\
 | \bar{A}(m-1), \bar{L}(m), X > m]
 \end{aligned}$$

for all possible values of  $a(m)$  in support of  $A(m)$ . One possibility is to specify and fit an appropriate multivariate regression model with the possible values of  $a(m)$  indexing the multivariate outcomes at time  $m$ .

To understand why this is the correct algorithm, we first note that any regimen at  $m$  can be a function of  $X$  only if  $X \leq m$ , so that  $X$  is known by  $m$ . When  $X > m$ , we must average over  $X^{0(m)}_{\mathcal{G}_{\text{opt}(\beta, \psi), m+1}}(\psi)$  because  $X^{0(m)}_{\mathcal{G}_{\text{opt}(\beta, \psi), m+1}}(\psi)$  is a function of  $X$ . When  $X > m$ ,  $X^{a(m)}_{\mathcal{G}_{\text{opt}(\beta, \psi), m+1}}(\psi)$  will be the value of  $X^{\mathcal{G}_{\text{opt}(\beta, \psi), m}}(\psi)$  if the optimal regimen  $\mathcal{G}_{\text{opt}(\beta, \psi)}$  dictates the exposure  $a(m)$ . The optimal regimen will choose the  $a(m)$  that optimizes the contribution to the utility at time  $m$ . But the optimizing  $a(m)$  depends on the  $a(k)$  chosen for the regimen for  $k > m$ . Thus, we need to use backward recursion to estimate the optimal regimen.

To be more specific, consider the subgroup of subjects with a history  $(\bar{A}(K-1), \bar{L}(K), X)$  with  $X < K$  so  $X = X^{0(K)}_{\mathcal{G}_{\text{opt}, K+1}}(\psi^*) \in \bar{L}(K)$ . Then  $a(K) = \mathcal{G}_{\text{opt}(\beta, \psi), K}^*[\bar{A}(K-1), \bar{L}(K)]$  that maximizes  $\gamma_K[a(K), \bar{A}(K-1), \bar{L}(K), X, \beta^*]$  is the optimal treatment choice at  $K$ . However, we are only considering regimens (interventions) that do not force subjects to gain weight. We now argue that for any subject with  $A(K)$  less than  $\mathcal{G}_{\text{opt}(\beta, \psi), K}^*[\bar{A}(K-1), \bar{L}(K)]$ , the optimal decision is not to intervene at all, so the subject receives his observed treatment  $A(K)$ . The subject with  $A(K)$  less than  $\mathcal{G}_{\text{opt}(\beta, \psi), K}^*[\bar{A}(K-1), \bar{L}(K)]$  could still have received any treatment between 0 and  $A(K)$ . However, among these set of treatments, the treatment  $A(K)$  is optimal by the concavity condition above.

Next consider the subgroup of subjects with a history  $(\bar{A}(K-1), \bar{L}(K), X)$  with  $X > K$ . To find the optimal treatment, we average over  $X^{0(K)}_{\mathcal{G}_{\text{opt}, K+1}}(\psi^*)$ . As the average over  $X^{0(K)}_{\mathcal{G}_{\text{opt}, K+1}}(\psi^*)$  of a function that is concave in  $a(K)$  for every possible value of  $X^{0(K)}_{\mathcal{G}_{\text{opt}, K+1}}(\psi^*)$  remains a concave function of  $a(K)$ , we again take  $\mathcal{G}_{\text{opt}(\beta, \psi), K}[\bar{A}(K-1), \bar{L}(K)] = \min\{A(K), \mathcal{G}_{\text{opt}(\beta, \psi), K}^*[\bar{A}(K-1), \bar{L}(K)]\}$ .

That the same argument holds for each  $m$  is a standard dynamic programming argument as discussed in Robins.<sup>8</sup>

As  $(\beta^*, \psi^*)$  are unknown we must estimate them by  $g$ -estimation. Define

$$X_m^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\psi) = X_{\mathcal{G}_{\text{opt}(\beta, \psi), m}}^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\psi)$$

$$Y_m^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\beta, \psi) = Y - \sum_{j=m}^K \gamma_m[A(j), \bar{A}(j-1), \bar{L}(j), X_j^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\psi), \beta].$$

Note these equations are much more complex than the equations for  $\psi$  using an SNFTM and SNMM for a fixed  $g$  in that  $\mathcal{G}_{\text{opt}}$  is now not known but depends on the parameters  $(\beta, \psi)$  through the above algorithm for  $\mathcal{G}_{\text{opt}(\beta, \psi)}$ . Thus, we can no longer estimate  $\psi^*$  independently of  $\beta^*$  as  $X_m^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\psi)$  is now a function of  $\beta$  as well as  $\psi$  through its dependence on  $\mathcal{G}_{\text{opt}(\beta, \psi)}$ . Rather, we must solve both pairs of  $g$ -estimation equations simultaneously.

Specifically, given the optimal regimen RC and CD assumptions, to obtain CAN estimators of the unknown parameters, we find jointly  $(\tilde{\beta}, \tilde{\psi})$  so that both the score test for the covariate vector depending on  $X_m^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\psi)$  is precisely zero and the score test for the covariate vector depending on  $Y_m^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\tilde{\beta}, \tilde{\psi})$  is precisely zero (both tests are restricted to subjects with  $X_m^{\mathcal{G}_{\text{opt}(\beta, \psi)}}(\psi) > \zeta$  and  $\Xi(m) = 1$ ). This turns out to be a very difficult computational problem. Robins<sup>8</sup> describes a number of computational simplifications, but they are beyond the scope of the current paper. Finally, we obtain  $\mathcal{G}_{\text{opt}((\tilde{\beta}, \tilde{\psi}))}$  as our estimate of the optimal regimen  $\mathcal{G}_{\text{opt}(\beta^*, \psi^*)}$  and  $n^{-1} \sum_i^n Y_{0i}^{\mathcal{G}_{\text{opt}(\tilde{\beta}, \tilde{\psi})}}[(\tilde{\beta}, \tilde{\psi})]$  as our estimate of the expected utility  $E[Y_0^{\mathcal{G}_{\text{opt}}}]$  under the optimal regimen.

Both estimation of  $E[Y_0^g]$  for a known  $g$  and of  $E[Y_0^{\mathcal{G}_{\text{opt}}}]$  can be modified to allow for censoring at the end of follow-up at  $K+1$  and for intractable unmeasured confounding in certain subgroups using methods exactly analogous to the methods for the estimation of  $E[Y_0]$ .