University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 140

Comparison of the Inverse Probability of Treatment Weighted (IPTW) Estimator With a Naïve Estimator in the Analysis of Longitudinal Data With Time-Dependent Confounding: A Simulation Study

Thaddeus Haight* Romain Neugebauer†
Ira B. Tager‡ Mark J. van der Laan**

*Division of Epidemiology, School of Public Health, University of California, Berkeley, tad@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, romain.s.neugebauer@kp.org

‡Division of Epidemiology, School of Public Health, University of California, Berkeley, ibt@berkeley.edu

**Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper140

Comparison of the Inverse Probability of Treatment Weighted (IPTW) Estimator With a Naïve Estimator in the Analysis of Longitudinal Data With Time-Dependent Confounding: A Simulation Study

Thaddeus Haight, Romain Neugebauer, Ira B. Tager, and Mark J. van der Laan

Abstract

A simulation study was conducted to compare estimates from a naïve estimator, using standard conditional regression, and an IPTW (Inverse Probability of Treatment Weighted) estimator, to true causal parameters for a given MSM (Marginal Structural Model). The study was extracted from a larger epidemiological study (Longitudinal Study of Effects of Physical Activity and Body Composition on Functional Limitation in the Elderly, by Tager et. al [accepted, Epidemiology, September 2003]), which examined the causal effects of physical activity and body composition on functional limitation. The simulation emulated the larger study in terms of the exposure and outcome variables of interest– physical activity (LTPA), body composition (LNFAT), and physical limitation (PF), but used one time-dependent confounder (HEALTH) to illustrate the effects of estimating causal effects in the presence of time-dependent confounding. In addition to being a time-dependent confounder (i.e. predictor of exposure and outcome over time), HEALTH was also affected by past treatment. Under these conditions, naïve estimates are known to give biased estimates of the causal effects of interest (Robins, 2000). The true causal parameters for LNFAT (-0.61) and LTPA (-0.70) were obtained by assessing the log-odds of functional limitation for a 1-unit increase in LNFAT and participation in vigorous exercise in an ideal experiment in which the counterfactual outcomes were known for every possible combination of LNFAT and LTPA for each subject. Under conditions of moderate confounding, the IPTW estimates for LNFAT and LTPA were -0.62 and -0.94, respectively, versus

the naïve estimates of -0.78 and -0.80. For increased levels of confounding of the LNFAT and LTPA variables, the IPTW estimates were -0.60 and -1.28, respectively, and the naïve estimates were -0.85 and -0.87. The bias of the IPTW estimates, particularly under increased levels of confounding, was explored and linked to violation of particular assumptions regarding the IPTW estimation of causal parameters for the MSM.

INTRODUCTION

In a paper entitled, *Longitudinal Study of Effects of Physical Activity and Body Composition on Functional Limitation in the Elderly,* by Tager *et. al*. (accepted, Epidemiology, September 2003), we applied methods for estimating parameters of interest defined by Marginal Structural Models (MSMs) to an observational, longitudinal study of the causal effects of physical activity and body composition on physical functioning (the self-reported ability to carry out defined physical tasks) in the elderly. After review of the manuscript, we were asked by the editor of the journal (Epidemiology) to apply these methods for MSMs using a simple example, and demonstrate their advantages compared to standard conditional analytical methods. We chose to use MSMs for two reasons: 1) our observational data demonstrated time-dependent confounding which can lead to bias in the estimates from standard conditional analysis (e.g., unweighted logistic or multiple linear regression); and 2) we were interested in estimates of the population (marginal) effects of physical activity and body composition.

To comply with the editor's request, we conducted a simple simulation to emulate the original study that we carried out insofar as the main question of interest was concerned (i.e., the causal effects of physical activity and body composition on subsequent physical functioning). The purpose of the simulation is to assist biostatisticians and epidemiologists to understand: 1) the purpose and conceptual basis of the application of MSMs; 2) the estimation of these models; and 3) the interpretation of the results from MSMs and how these results compare to those from standard approaches. However, for simplicity and ease of presentation, we only used a subset of the variables and time points from the original analysis. Again for simplicity, we assumed that all

subjects were followed until the end of the study, unlike the larger study in which censoring occurred. Neither the restrictions imposed on the number of variables nor the assumption of no censoring affected the interpretations derived from this simulation.

PURPOSE OF MSMs AND ESTIMATION OF CAUSAL EFFECTS

MSMs are used to define causal parameters of interest in exposure-response relationships based on the notion of counterfactuals (Robins, 2000). This notion essentially allows us to assess observational data in a hypothetical world in which, contrary to fact, the subjects were exposed to all possible levels of an exposure, and had outcomes associated with all those possible exposures (i.e. each subject serves as his/her own control and the outcomes are known for every exposure they were exposed to). With counterfactual data in hand, a researcher can evaluate whether differences in the outcome are causally attributable to differences in the level of the exposure, using an appropriate estimator of the causal parameters of interest defined by a MSM. In our case, we applied the notion of counterfactuals and assumed a MSM for examining whether differences in physical function (i.e. mean differences) were linked causally to differences in levels of physical activity and body composition. Given the presence of time-dependent confounding in our data, we used the Inverse Probability of Treatment Weighted Estimator (IPTW) (Robins, 1999) to obtain asymptotically unbiased estimates of the causal exposure-response relationship. In the presence of time-dependent confounding, unbiased estimates of these causal parameters cannot be obtained from standard analytical approaches.

DESCRIPTION OF OBSERVED SAMPLE DATA AND FULL POPULATION DATA USED IN THE SIMULATION

We simulated observed sample data for n=1000 female subjects for two time points. These observed data represent sample data that could be drawn from a population of elderly women, and, for purposes of this simulation, we know in advance the underlying distributions of the variables for the population. The observed data vector for a given subject over the two time points (*j=0,1*) is defined as :

$$O=(Age\ ,L(0),A_1(0),A_2(0),Y(0),L(1),A_1(1),A_2(1),Y(1))$$

where the exposures of interest are $A_1(j)$ that represents physical activity (LTPA, 1=engage in vigorous activity/ 0=do not engage in vigorous activity), and $A_2(j)$ that represents lean:fat ratio (LNFAT, body composition measured as a continuous variable in the form of a ratio of lean body mass to fat mass). We also define covariates *Age* (continuous variable of age at entry into the study, time *j=0), L(j)* as health status (1=medical condition/0=no medical condition), and the outcome measure *Y(j)* as physical function (1=impaired/0=not impaired). All of these variables are time-dependent over the two time points that we assess these subjects *(j=0,1)*, except age which is fixed at its baseline value (*j=0*). Both Age and LNFAT are centered on their respective median values for purposes of interpretation. We are interested primarily in the causal effects of both body composition (LNFAT) and physical activity (LTPA) on physical function (PF). However, given that health status (HEALTH) is a time-dependent confounder of the relationship of interest, we can not obtain unbiased causal estimates of the relationship of interest with standard analytical methods.

We define the full population data, referred to hereafter as the "full" data, as representing the collection of all counterfactual variables for the entire population of

elderly women from which the observed data above were sampled. In other words, we define this full data for a hypothetical world in which, contrary to fact, a subject was exposed to all possible levels of LTPA and LNFAT and had PF "counterfactual" outcomes associated with all those possible exposures. If we had access to the full data, we could estimate directly and conveniently the unconfounded causal effects of interest. For example, if we were interested in the causal effect of vigorous exercise on the odds of functional impairment, we could calculate the odds of impairment under the condition where everyone received a "treatment" of vigorous exercise and then compare this to the odds of impairment under the condition where everyone received a "treatment" of non-vigorous exercise. In observational studies, such full data are not observed, and we have to operate with data in which only one treatment regimen and the corresponding outcome are observed for each subject. Methods for estimating parameters of MSMs rely, in particular, on the assumption that the observed data are one component of the full data, with the remainder of the full data as missing, and provide an unbiased estimation of the parameters that would be obtained under the ideal circumstance of having the full data in hand. To evaluate how well the IPTW method retrieves causal parameters from observed data, we simulated observed data and full data, and compared the estimates we obtained from the observed data based on the IPTW with those from the full data.

SIMULATION PROCEDURE FOR GENERATION OF THE OBSERVED DATA AND FULL DATASETS (See Figure 1)

All of the procedures used to generate data and implement the IPTW estimator were performed with SAS Version 8.2 (See Appendix 2 for SAS Code). The observed data for N=1000 subjects used known data-generating distributions and parameter values based on those from the female data in the larger study by Tager *et. al.*

(a) Generate Age as a normally-distributed *N(70,64)* variable with the lowest value truncated at 50.

(b) Generate baseline Health L(0) using a Bernoulli distribution with data-specified probability of a medical condition

$$logitP(L(0)=1 \mid Age=age) = -0.20 + 0.05 * Age$$

(c) Generate baseline LTPA using a Bernoulli distribution with data-specified probability of vigorous exercise

$$logitP(A_1(0)=1 \mid Age=age, L(0)=l(0)) = 0.5 - 0.07 * Age - 0.70 * L(0)$$

(d) Generate baseline LNFAT using a normal distribution with data-specified mean

$$(1.5 + 0.006 * Age + 0.11 * A_1(0) - 0.07 * L(0)) \text{ and } \sigma^2 = 0.15$$

(e) Generate baseline PF using a Bernoulli-distribution with data-specified probability of impairment

$$\log itP(Y(0)=1 \mid Age=age, L(0)=l(0), A_1(0)=a_1(0), A_2(0)=a_2(0)) =$$

$$-1.00 + 0.08 * Age + 0.40 * L(0) - 1.20 * A_1(0) - 0.80 * A_2(0)$$

(f) Generate Health at follow-up using a Bernoulli-distribution with data-specified probability

$$\log itP(L(1)=1 \mid Age=age, A_1(0)=a_1(0), A_2(0)=a_2(0), Y(0)=y(0))$$

$$0.20 + 0.05 * Age - 0.80 * A_1(0) - 0.20 * A_2(0) + 0.20 * Y(0)$$

(g) Generate follow-up LTPA using a Bernoulli-distribution with data-specified probability

$$\log itP(A_1(1)=1 \mid Age=age, A_1(0)=a_1(0), L(1)=l(1), Y(0)=y(0))$$

$$-1.20 - 0.07 * Age + 1.80 * A_1(0) - 0.30 * L(1) - 0.70 * Y(0)$$

(h) Generate follow-up LNFAT using a normal distribution with data-specified mean

$(1.45 + 0.002 * Age + 0.012 * A_1(1) + 0.04 * L(1) + 0.46 * A_2(0) - 0.02 * Y(0))$,

and variance $\sigma^2 = 0.04$

(i) Generate follow-up PF using a Bernoulli-distribution with data-specified probability

$$\log it P(Y(1) = 1 \mid Age = age, A_1(1) = a_1(1), A_2(1) = a_2(1), L(1) = l(1), Y(0) = y(0))$$

$$-1.7 + 0.02 * Age - 0.50 * A_1(1) - 0.40 * A_2(1) + 0.50 * L(1) + 2.30 * Y(0)$$

To simulate a full dataset that represents the counterfactual data that could have been collected for the entire population of elderly women, we generated a sample of N=10,000 subjects and followed the procedure above for parts (a) and (b) to create a population of subjects with the same underlying probability distributions of baseline age and health as our observed group. In addition, we substituted the following procedure for steps c-i above: for each subject, we fixed 42 unique combinations of LTPA ($A_1(j)=0,1$) and LNFAT ($A_2(j)=0.5,\ldots 2.5$ *by 0.1 increments*) to simulate the data from an ideal experiment in which a subject experienced all possible exposures at time $j=0,1$. The simulation incorporated all possible combinations of LTPA and LNFAT over time, based on the observed data. For example, the difference in LNFAT over time did not exceed 0.8 for biological reasons, so the full data were structured in such a way as to allow only combinations of LNFAT over time that did not exceed 0.8. All combinations of LTPA over time were included in the simulation. The counterfactual outcomes $Y_{\bar{a}}(0), Y_{\bar{a}}(1)$ [1], as well as health at follow-up $L_{\bar{a}}(1)$ [1], were generated using steps e, i, and f, respectively.

[1] $\bar{a}(0) = (a_1(0), a_2(0))$, $\bar{a}(1) = (a_1(0), a_2(0), a_1(1), a_2(1))$

(See Appendix 2 for details). In essence, these counterfactual variables represent all the possible outcomes and covariates associated with all possible histories of LTPA and LNFAT experienced by any given subject over time in the ideal experiment. Again, to emphasize the differences in the simulated data, we can view the full data as the data for a population of subjects from which we sample our observed data and, based on that sample, we observe only one set of possible values for each subject at each time point.

IPTW ESTIMATION OF THE PARAMETERS OF THE MSM

A detailed treatment of the assumptions that underlie the application of MSMs to longitudinal data structures is provided by Robins (1999), Robins and van der Laan (2002), and Yu and van der Laan (2002a). In general terms, we assume: 1) the existence of counterfactuals and link the observed data to these counterfactuals; 2) temporal ordering of the data (i.e. the occurrence of covariate *L* and exposures *A* precede the outcome *Y* in the causal pathway, and the *L* influences the exposures *A*); 3) no unmeasured confounding (sequential randomization) of the exposures *A(t)*. In addition, we assume all *A(t)* are possible for given covariate history *L(t)*, conditional on history of *A* up to time *t* (Experimental Treatment Assignment, ETA). The theory and practical implications of this assumption are explored by Neugebauer and van der Laan (2003).

In this simulation, we assume the MSM:

$$\text{logit}\Pr(Y_{\bar{a}}(t)=1) = \beta_0 + \beta_1 * a_1(t) + \beta_2 * a_2(t)$$

This model states that the logit of the counterfactual outcome $Y_{\bar{a}}$ (PF), indexed by treatment regime $\bar{a}=(\bar{a}_1, \bar{a}_2)$, can be modeled by a linear combination of the effect of a_1 (LTPA) and a_2 (LNFAT) at time *t*. It is very likely that we have misspecified the true MSM. Other model choices could have been made to specify the MSM; however, we

assume the given model because it is appropriate to answer the research question of interest.

The β's can be estimated by solution of the IPTW estimating equation. Solution of this estimating equation is equivalent to performance of a weighted regression of *Y(t)* on $A_1(t)$ and $A_2(t)$ with the use of subject-specific weights. Under the sequential randomization assumption (SRA), the denominator of these weights corresponds in practice with an estimate of the probability of observing a subject's actual history of exposure up to *t* given their past history. Here we use stabilized weights to reduce the variability in the estimation of the causal parameters (Robins, 2000). Any weights less than 0.2 or greater than 5 were truncated at these values. These subject-specific IPTW weights are estimated based on the observed data using what is termed as a treatment model that includes all the possible confounders of A_1 *and* A_2:

$$Sw(t) = \prod_{k=0}^{t} (Pr(A_1(k) \mid \overline{A}_1(k-1))/(Pr(A_1(k) \mid \overline{A}_1(k-1), \overline{Y}(k-1), \overline{L}(k), Age) * \quad [1]$$
$$(Pr(A_2(k) \mid A_1(k), \overline{A_2}(k-1)))/(Pr(A_2(k) \mid A_1(k), \overline{A_2}(k-1), \overline{Y}(k-1), \overline{L}(k), Age)$$

where the exposures, LTPA and LNFAT, are indexed by 1 and 2 (i.e., separate weights are estimated for each exposure). In the case of LNFAT, which is a continuous variable, we estimate the conditional density of LNFAT, given past LNFAT, LTPA and covariates with linear regression. We assume the error is normally distributed in the LNFAT model. In the case of LTPA, which is a binary variable, we estimate the conditional probabilities of the physical activity categories by logistic regression. We refer occasionally to the results associated with the IPTW estimator as "weighted", because the estimator weights

[1] When k=0, $\overline{Y}(k-1)$ and $\overline{A}(k-1)$ empty sets.

the observed data in the regression by the products of the subject-specific weights up to time *t* that we can calculate directly based on the equation above. Weighting the observed data in the regression has the effect of "correcting" for the confounding in the data that interferes with the causal estimation of the exposure-response relationship. By comparison, we conducted a standard, "naïve" analysis which carried out the same regression without weights. We will refer to the estimates from this naïve estimator as "unweighted" and/or "naïve".

We examined the effects of different levels of confounding of LTPA and LNFAT to illustrate these effects on the model parameter estimates (See Appendix 1 for Details).

An unweighted regression of $Y_{\bar{a}}(t)$ on $a_1(t)$ and $a_2(t)$ was performed using the full population dataset. The parameter estimates from this regression can be interpreted as the "true" causal parameters representing the causal effect of LNFAT and LTPA on PF, for the MSM that we have assumed.

All regressions were implemented with general estimating equations (GEE) with the weighting option SCWGT (PROC GENMOD of SAS v8.2). Given that the standard errors (SEs) from the weighted regression are conservative (i.e. larger than the true variance), we used a bootstrap to estimate the SEs empirically.

RESULTS

The distributions of the variables and the level of confounding of LTPA and LNFAT are comparable to the same subset of variables in the female data from the original study (data not shown) on which we based this simulation (Tables 1-2). We would expect the level of confounding in the original study to be greater and more complex compared to the level of confounding in this simulation, given that we investigated eight confounders of LNFAT and LTPA (several of which were time-

dependent confounders) in the original study compared with two in this simulation. The relationship between health at follow-up and past exposure (LNFAT, LTPA) and outcome (PF) (Table 3) clearly illustrates the importance of the health variable as a nuisance factor on the causal pathway between past exposures and the outcome. If we examined the effects of LNFAT and LTPA on PF in a standard conditional model that adjusted for health, we would violate the assumption that health is held constant in the model, since a change in either LNFAT or LTPA would likely produce a change in health. The implications of such a violation are that the estimates of the causal effects of LNFAT and LTPA would be biased and uninterpretable.

Table 4 summarizes the findings of this simulation. The differences in the parameter estimates for the MSM that we specify above reflect: 1) the different estimators being used to estimate the model parameters; and 2) the impact of the different confounding structures that were simulated in the data (i.e. confounding in and of itself can be considered a nuisance parameter when trying to estimate parameters of interest). The distributions of the IPTW weights (medians, ranges) are based on calculations of the *Sw(t)* equation in the Methods Section. Percentages of truncated values for some of the different sets of weights are included in the footnote at the bottom of Table 4. Based on the distribution of the weights, the level of confounding in the data can be assessed to some extent by the deviation from 1 in the weights (the greater the deviation from 1, the greater the confounding).

Estimates from the regression based on the full population dataset are in the bottom row of the table. These estimates represent the best approximation (SE = 0.001) of the true, unconfounded, causal estimates of the effects of LNFAT and LTPA on PF, given our assumed MSM.

We find that the parameter estimates from the assumed MSM (Table 4, column labeled IPTW), compared to the naïve estimates from the unweighted models, are generally closer true parameter values of the full data for the intercept and LNFAT but not for LTPA. We interpret the LNFAT coefficient in the MSM as the causal effect if, contrary to fact, the entire population experienced a one unit increase in LNFAT over the observed population median value. A similar interpretation is given to the LTPA coefficient and to the sum of the effects of the two treatments together. The intercepts represent the reduction in log-odds of future functional limitation if, contrary to fact, the entire population was at the median level of LNFAT and not exercising vigorously.

Direct comparisons between the parameter estimates from the MSM and the conditional models without weights are not appropriate, since the former provides population (marginal) estimates and the latter provides conditional (i.e., stratum-specific) estimates. However, we make this comparison for heuristic purposes, since epidemiologists often attach marginal inferences to these conditional estimates.

At levels of confounding that approximate those in the complete study data, the unweighted analysis that excludes the time-dependent confounder, "health" (Table 4, column labeled "Age and Health Omitted"-upper 3 rows), provides parameter estimates that are similar to those estimated in the MSM and by the full data. When age and the time-dependent confounder, "health" are added into the unweighted analysis (Table 4, column labeled "All Covariates"-upper 3 rows), the parameter estimates differ substantially from those of the MSM and from the "true" parameters for the full data. When a data set is simulated with levels of confounding that are greater than those in the actual study data, the intercept and parameter estimate for LNFAT of the MSM are close to those from the full population data. In contrast, estimates based on the unweighted analyses without age and health, differ substantially from those of the assumed MSM.

When age and health are added, the parameter estimates for the intercept and LNFAT are far from those of the MSM and the "true" parameters. However, the estimate for LTPA is closer to the "true" parameter than the estimate from the assumed MSM. Throughout these results, the MSM estimates have larger variances than the estimates from the unweighted analyses. If one treats estimates from the unweighted analysis as marginal estimates, the weighted and unweighted analyses would lead to considerable differences in the estimates of the population-level effects, with those from the unweighted analysis most likely to be biased or more severely biased than those obtained from the assumed MSM.

DISCUSSION

Methods for estimating causal effects defined by MSMs should be considered as an alternative set of analytical tools for estimation of exposure-response relationships in lieu of standard analytical approaches for two important reasons:

1. Longitudinal data that are encountered in practice often will be confounded in a time-dependent manner to some degree. Depending on the level of confounding exhibited by the data, estimates to measure exposure-response relationships could vary widely from estimates in which relationships are unconfounded. Epidemiologists often assume that adjustment for a set of variables that confound a specific exposure-response relationship provides an estimate of the direct, unconfounded effect of interest of the exposure on the response. However, adjustment for these suspected confounding variables, which in many cases is necessary, often leads to the inclusion of variables that actually are on the causal pathway between past exposure variables and the response-- a circumstance that can lead to biased effect estimates. Moreover, the time-dependent nature of confounding usually is

ignored, where the exposure-response relationship is assessed over time, which also contributes to bias. The methodology that we have illustrated in this simulation addresses the problems of the estimation of effects under these circumstances and provides estimates that are less biased with respect to those estimates that we would obtain given truly unconfounded data.

2. Another reason to consider these methods is that they provide estimates that can be of more practical value from a public health perspective. In this context, estimates of the marginal effects of different levels of exercise on reductions in the risk (measured as the log odds in this simulation) of functional impairment are of interest, rather than estimation of the conditional (stratum-specific) effect of exercise that is obtained with standard statistical methods. From a population perspective, the interpretation of conditional effects can be limited by the fact that: 1) the factors on which we condition are a set of nuisance variables that are not of primary interest to a given study; and 2) in the presence of time-dependent confounding, the strata defined by the adjustment will change for a given change in the level of the exposure variable. Moreover, the methods used for estimating parameters of MSMs allow for estimation of stratum-specific, marginal effects, when such inferences are of interest (e.g., the identification of susceptible sub-groups within a population) (Robins, 2000). MSMs also can be specified in ways that examine the interaction effects of multiple exposures ("treatments") with respect to an outcome of interest (Robins, 1999).

Based on the results of this simulation, the IPTW estimates are generally less biased than the unweighted estimates (i.e. standard regression estimates) when both sets of estimates are compared to the "true", unconfounded parameter values for the full data. However, the unweighted estimate for LTPA, adjusted for age and health, is closer to the "true" parameter value for LTPA. The bias of the IPTW estimate of LTPA, which is

particularly obvious under increased levels of confounding (IPTW estimate, -1.28), has three plausible sources[1]. One is a violation of the ETA assumption. Under this assumption, all possible treatments can occur for covariate history, conditional on past treatment. We exaggerated the level confounding of LTPA through covariate history (Appendix 1). This action had the effect of "predetermining" treatment (LTPA), given covariate history, which led consequently to a biased IPTW estimate of LTPA. In Figures 2(a) and 2(b), this effect is shown by the clustering of predicted probabilities of less than 0.1 and greater than 0.9 near the observed values of 0 and 1, respectively. By comparison, we increased the level of confounding of LNFAT through past LNFAT rather than covariate history. The IPTW estimate for LNFAT was not as biased because beyond the effects of conditioning on past LNFAT, covariate history did not "predetermine" treatment (LNFAT) (data not shown).

Another source of bias of the IPTW estimate could have resulted from the misspecification of the assumed MSM itself. Under moderate levels of confounding (Table 4, Box 1), bias occurs (IPTW estimate-0.94 vs "true" LTPA –0.70), but it is not as dramatic as when the level of confounding is increased (-1.28 vs –0.70). We might conjecture that the additional confounding appears to place a greater demand on the correct specification of the MSM, because the bias of the IPTW estimate increases. If we had specified LTPA differently in the MSM, we may have seen IPTW estimates that were closer to the "true" LTPA parameter. LNFAT may not have been misspecified in the MSM, at least not as severely, given that the IPTW estimates for LNFAT are consistently close to the "true" LNFAT parameter value based on the full data.

[1] A technical discussion with regard to the violation of the ETA assumption and misspecification of MSMs is found in Neugebauer R, van der Laan, M. Locally efficient estimation of non-parametric causal effects on mean outcomes in longitudinal studies (July 25,2003) at {http://bepress.com/ucbbiostat/}.

Lastly, the use of stabilized (Sw_i) vs. non-stabilized (w_i) weights can lead to additional bias in the IPTW estimator in instances where the MSM has been misspecified. Figure 3 illustrates a hypothetical situtation in which the IPTW estimates of a causal effect of interest are different based on using a misspecified MSM with stabilized and non-stabilized weights, $m(a|\beta_{Sw})$, $m(a|\beta_W)$, represented by the lines A and C in the figure (the IPTW estimates, β_{Sw}, β_W, are the slopes of these lines). For comparison, the curvilinear line, line B, represents the correctly specified MSM of the causal effect of interest $m(a|\beta)$. The stabilized weights used for the IPTW estimator do not aim at the true parameters defined by the full data; that is, the IPTW estimates we reported, using stabilized weights, provide an approximation of the real causal effect, but it is an approximation of this causal effect for the levels of LTPA/LNFAT that are the most represented in the observed data. For example, if the observed data are mainly represented by the first two levels of x ($x=1,2$) in figure 3, then by using stabilized weights, we give more weight to those levels in the regression, and we would obtain an IPTW estimate β_{Sw} that approximates the slope of line A. On the contrary, the full data estimates we reported give an overall approximation of the causal effect, not so much an approximation of the effect for the levels of LTPA/LNFAT that are the most represented in the observed data. The non-stabilized weights, rather than the stabilized weights, used for the IPTW estimator aim at the true parameters defined by the full data. In other words, additional weight would not be given to the first two levels of x, and we would obtain an IPTW estimate β_W that instead approximates the slope of line C in figure 3. Whether stabilized weights or non-stabilized weights are used for the IPTW estimator is arbitrary and left to the discretion of the investigator. Ultimately, he/she will decide if the IPTW estimator should be more data-driven and the causal effects of interest should depend or not depend on the levels of exposure (treatment) that occur more

predominantly in the data. Here we have simply illustrated the bias that occurs when the assumptions behind the IPTW estimator being used do not reflect what is being ascertained with the full data. A correctly specified MSM, in any event, will yield the same IPTW estimates for stabilized or non-stabilized weights.

We have demonstrated that the implementation of the IPTW estimator is straightforward with standard statistical software. However, we also have demonstrated, through various violations of the assumptions that underlie the implementation of the IPTW estimator, that the estimator can be biased. We clearly showed this when we exaggerated the level of confounding of LTPA through past covariate history, which resulted in a direct violation of the ETA assumption. With any implementation of the IPTW estimator, this assumption can be checked easily as we have shown with Figures 2(a) and 2(b). Other assumptions that underlie the IPTW estimator are not as easily assessed. We must assume there are no unmeasured confounders of the treatment variables of interest in these analyses through a correctly specified treatment model, and that we have considered fully an optimal specification of the treatment variables in the MSM. Adoption of these practices diminishes the possibility that the IPTW estimator will be biased (or at least reduces the bias), but does not eliminate the possibility of bias of the estimator altogether. Alternative estimators are available ("double robust" estimators), which are fully described and currently being used to protect against biased IPTW estimates (Robins and van der Laan, 2002), (Neugebauer and van der Laan, 2002), and (Yu and van der Laan, 2002). Although these estimators are not readily available in standard software packages, they can be implemented to yield "robust" IPTW estimates through either correct specification of the MSM or the treatment model. These same methods address the problem of violation of the ETA assumption, and result in more

efficient and unbiased (or less biased) MSM estimates than those that are obtained through the IPTW estimator.

Table 1. Univariate Distributions of Simulated Variables for Observed Data (n=1000)			
Variable		Time1 (t_1)	Time2 (t_2)
Baseline Age (t_1 only)	Mean (SD)	69.9 (8.1)	NA
Lean:Fat	Mean (SD)	1.52 (0.39)	1.47 (0.26)
% Vigorous Exercise		52.0	41.2
% Healthy		54.5	55.3
% Disabled		23.9	28.2

Table 2. Bivariate Distributions of Simulated Variables for Observed Data (n=1000)			
Variable 1	Variable 2	Time1 (t_1)	Time2 (t_2)
Healthy	Age (mean±SD)	69.0 (8.2)	68.4 (7.9)
	Lean:Fat (mean±SD)	1.55 (0.39)	1.46 (0.26)
	% Vigrorous Exercise	60.7	51.9
	% Disabled	18.9	21.3
Unhealthy	Age	71.1 (7.8)	71.8 (8.0)
	Lean:Fat	1.48 (0.38)	1.48 (0.26)
	% Vigrorous Exercise	41.5	28.0
	% Disabled	29.9	36.7
Vigorous Exercisers	Age	68.2 (8.0)	66.7 (7.4)
	Lean:Fat	1.58 (0.39)	1.49 (0.26)
	% Disabled	12.5	13.8
Not Vigorous Exercisers	Age	71.8 (7.8)	72.2 (7.9)
	Lean:Fat	1.46 (0.38)	1.46 (0.26)
	% Disabled	36.3	38.3
Disabled	Age	74.1 (7.4)	73.3 (7.8)
	Lean:Fat	1.43 (0.37)	1.45 (0.27)
Not Disabled	Age	68.6 (7.9)	68.6 (7.8)
	Lean:Fat	1.55 (0.39)	1.48 (0.27)

Table 3. Distributions of Age, Past Health, LNFAT, LTPA, and PF (t_1) by Health (t_2) in Observed Data (n=1000)		
Variable (t_1)	Healthy (t_2)	Unhealthy (t_2)
Age (mean±SD)	68.4 (7.9)	71.8 (8.0)
% Healthy	56.8	51.7
Lean:Fat (mean±SD)	1.55 (0.39)	1.48 (0.38)
% Vigrorous Exercise	63.7	37.6
% Disabled	18.4	30.7

Table 4. Parameter Estimates (standard errors)[1] based on IPTW and Naïve "Unweighted" Estimators.										
Level Of Confounding	**Estimates (SE)[2]**	**Naïve**		**IPTW**	**LTPA IPTW weights (untruncated)[3]**		**Lnfat IPTW weights (untruncated)[3]**		**LTPA x Lnfat IPTW weights (*Sw(t)*) (truncated 0.2,5)[4]**	
		Age and Health Omitted	**All Covariates**		**time1**	**time2**	**time1**	**time2**	**time1**	**time2**
1. Confounding of LTPA[5] and LNFAT consistent with that observed for these variables in the study data	Intercept	-0.68 (0.07)	-0.96 (0.09)	-0.68 (0.08)	0.91	0.87	0.98	0.97	0.93	0.85
	Lnfat	-0.57 (0.15)	-0.78 (0.16)	-0.62 (0.21)	(0.6– 3.2)	(0.4 – 5.6)	(0.5 – 2.5)	(0.3 – 3.1)	(0.5 – 4.4)	(0.21 – 5)[6]
	LTPA	-0.96 (0.11)	-0.80 (0.12)	-0.94 (0.13)						
2. Increased levels of confounding for LTPA and LNFAT relative to the study data	Intercept	-0.47 (0.07)	-0.96 (0.10)	-0.63 (0.09)	0.62	0.55	0.87	0.82	0.57	0.49
	Lnfat	-0.43 (0.12)	-0.85 (0.14)	-0.60 (0.16	(0.4 – 29)	(0.2 – 32)	(0.04–8.4)	(0.04- 9.7)	(0.2 - 5)[7]	(0.2 – 5)[8]
	LTPA	-1.51 (0.13)	-0.87 (0.16)	-1.28 (0.18						
3. Estimates from "full" data—"true" parameter values)	**Intercept -0.77 (0.005)** **LNFAT -0.61 (0.001)** **LTPA -0.70 (0.001)**									

[1] SEs based on standard deviations estimated for 1000 bootstrap IPTW and naïve estimates using original sample data.
[2] In unweighted models, GEE Estimates account for correlation (i.e. dependence) in response variable; in weighted models, independence of the response is assumed.
[3] Weights actually applied in the MSM regression are the truncated, stabilized weights Sw(t) that are described in the last column. Median (range) given for each set of weights.
[4] The distribution of the product of the treatment weight for females in the manuscript under review were: t_1: med, 0.89, range 0.2-5, with 1% of low weights truncated and 1% of high weights truncated: t_2: med 0.75, range 0.2-5.0 with 1% of low weights truncated and 1% of high weights truncated.
[5] Degree of confounding relates to confounding of the LTPA and LNFAT-outcome relationship by both AGE and HEALTH.
[6] 4 high weight values truncated at 5
[7] 1.3% of high weights and 4.1% of low weights truncated
[8] 1.5% of high weights and 9.2% of low weights truncated;

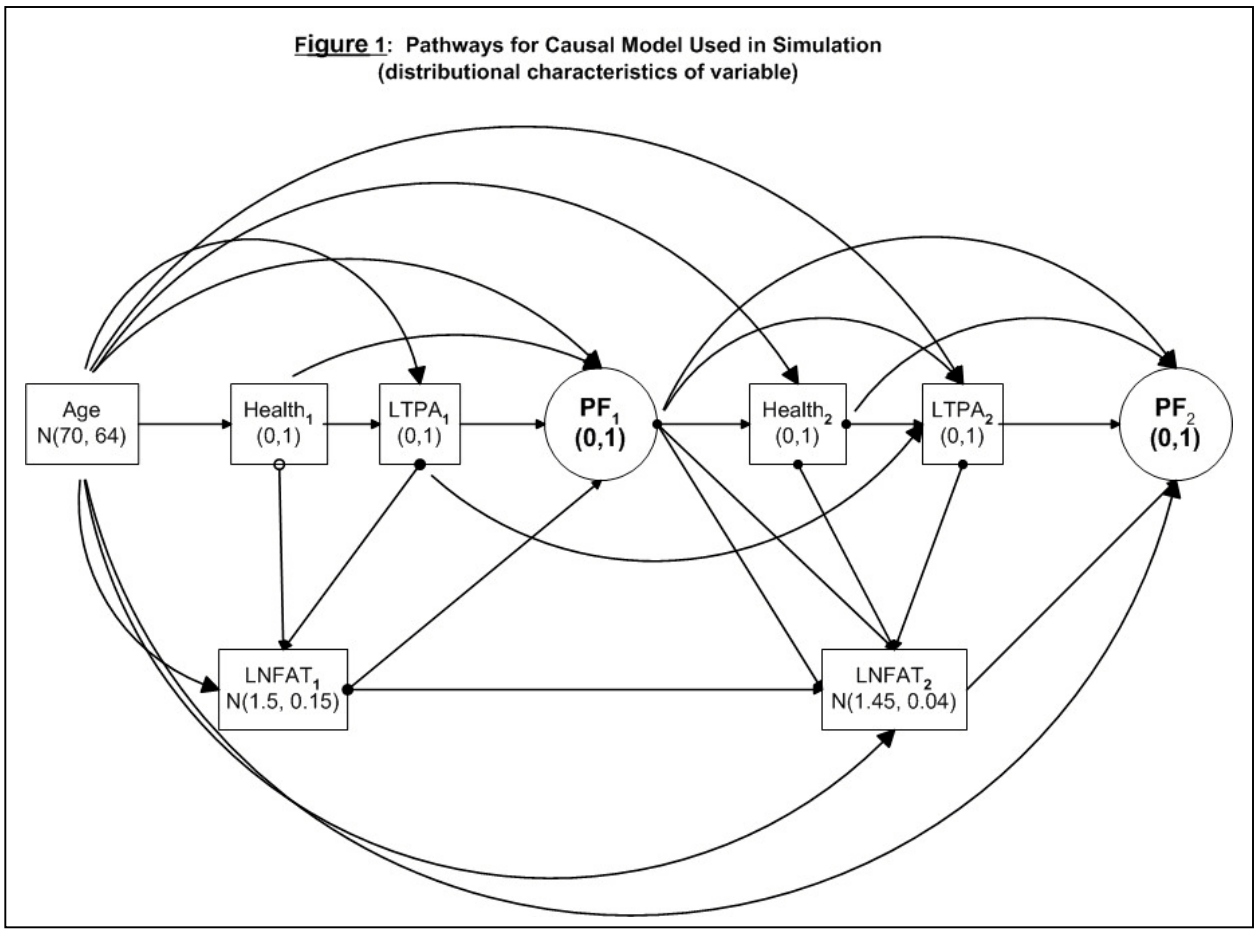


Figure 2a. Assessment of Experimental Treatment Assignment
Observed LTPA(0) vs. Predicted Probability LTPA(0)

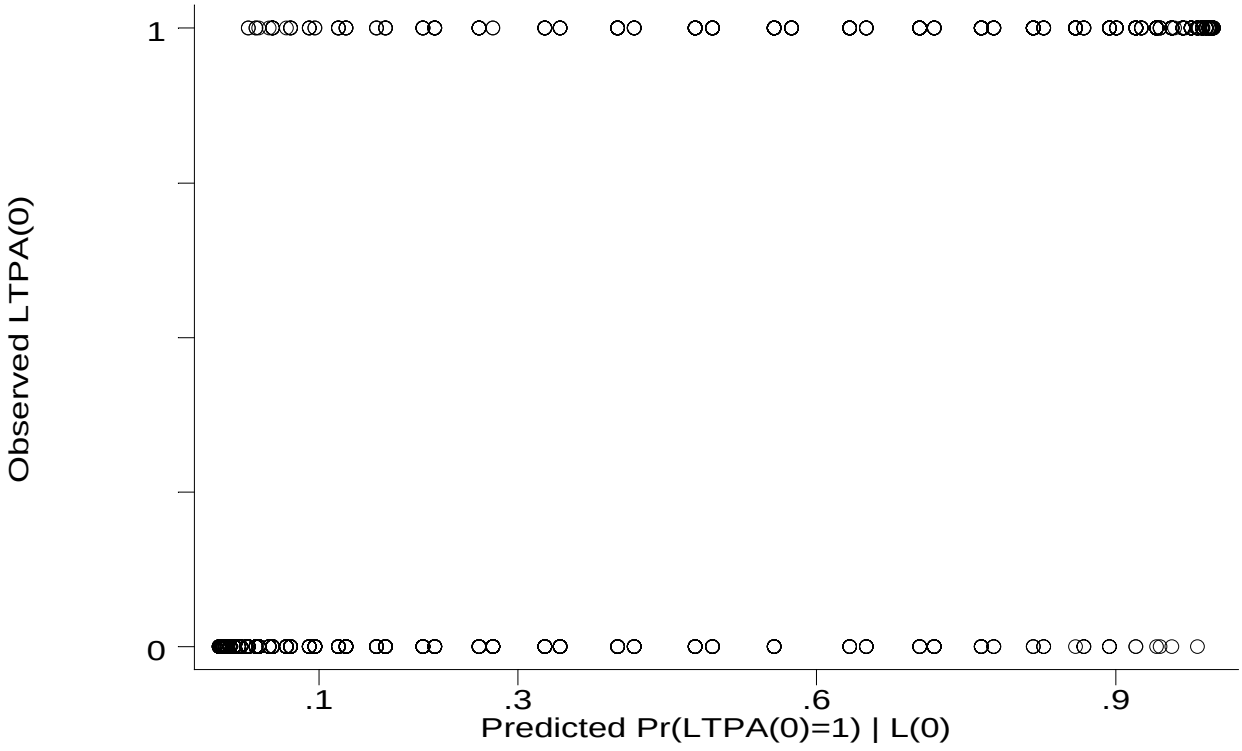


Figure 2b. Assessment of Experimental Treatment Assignment
Observed LTPA(1)-LTPA(0) vs. Predicted Probability LTPA(1)

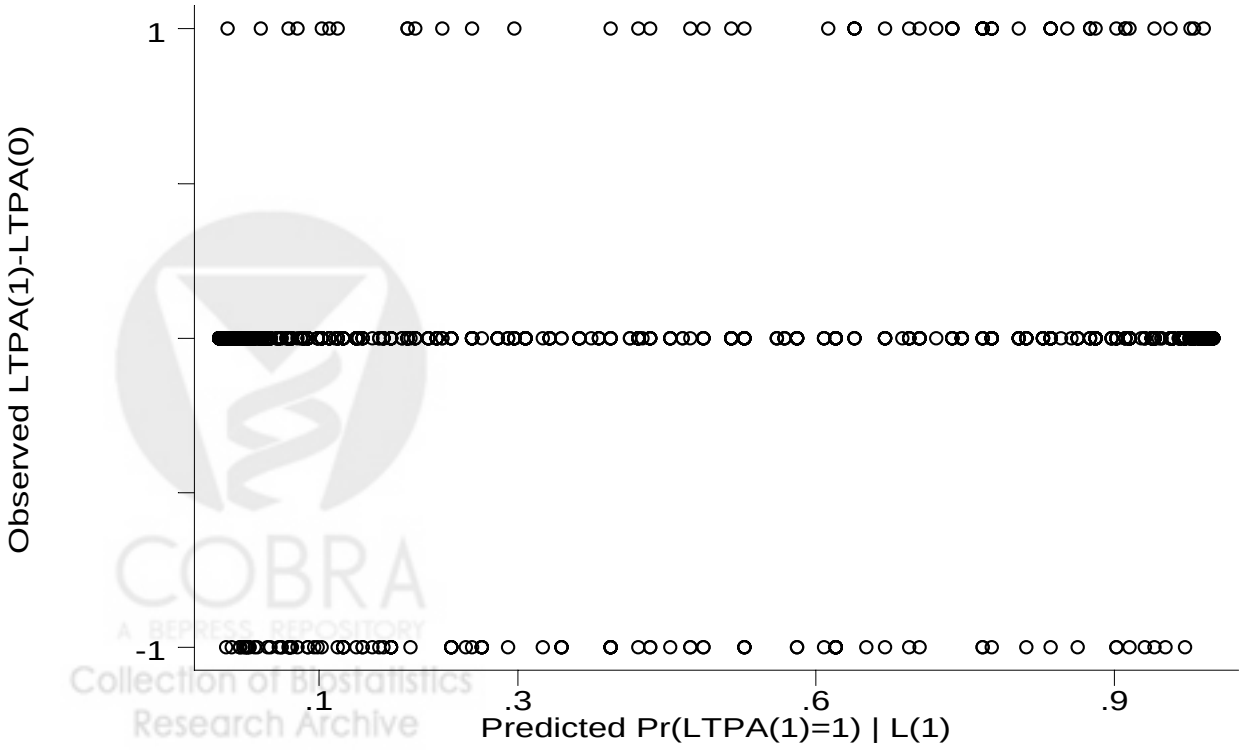
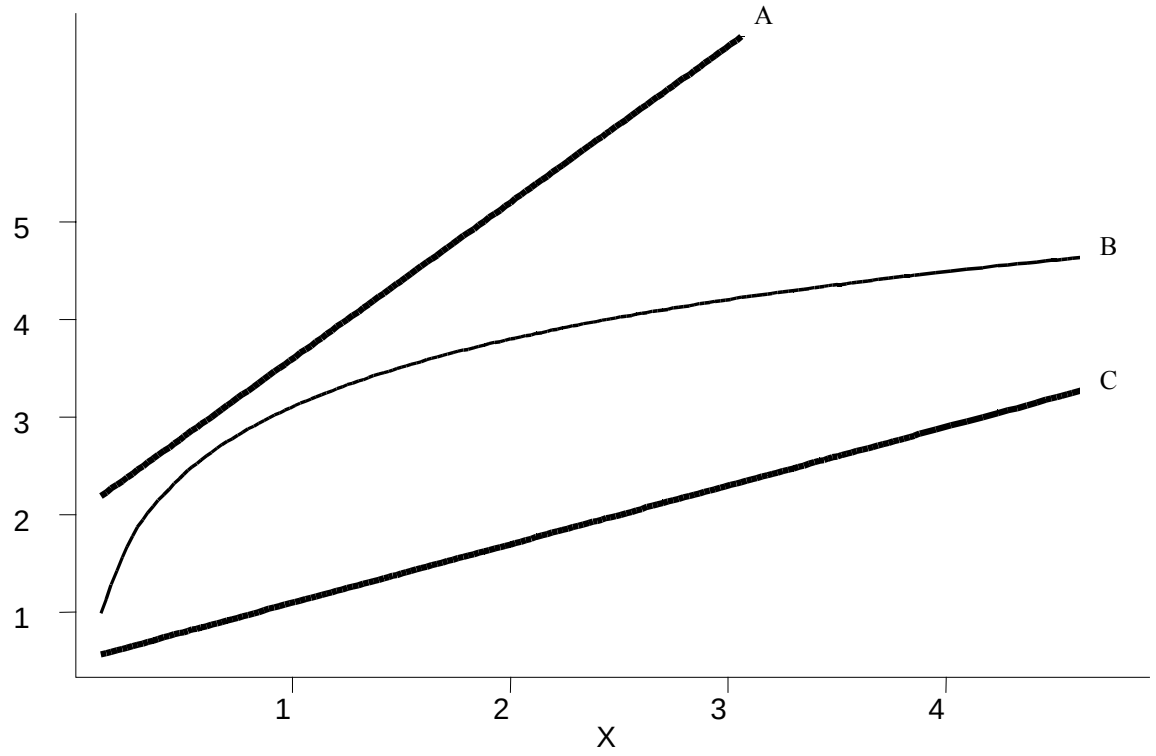


Figure 3. Comparison of IPTW Estimates when the MSM is misspecified



Assuming the observed data are mainly represented by levels $x=1,2$:
Line A represents a misspecified MSM using stabilized weights $m(a|\beta_{Sw})$ IPTW estimate $\beta_{Sw;}$
Line B represents a correctly specified MSM $m(a|\beta)$ of the causal effect of interest;
Line C represents a misspecified MSM using non-stabilized weights $m(a|\beta_{W})$ IPTW estimate $\beta_{W.}$

Appendix 1. – Data-Generation to increase level of confounding of LNFAT and LTPA.

All other variables in the simulation generated by the procedures given in the methods section. The sequence in the data–generation of all the variables in the simulation remained unchanged from the order shown in the Methods above.

(j) Generate baseline LTPA using a Bernoulli distribution with data-specified probability of vigorous exercise

$$\log itP(A_1(0)=1 \mid Age=age, L(0)=l(0)) = 0.5 - 0.3*Age - 1.8*L(0)$$

(k) Generate baseline LNFAT using a normal distribution with data-specified mean

$(1.5 + 0.03*Age + 0.11*A_1(0) - 0.2*L(0))$ and $\sigma^2 = 0.15$

(l) Generate follow-up LTPA using a Bernoulli-distribution with data-specified probability

$$\log itP(A_1(1)=1 \mid Age=age, A_1(0)=a_1(0), L(1)=l(1), Y(0)=y(0))$$

$$-1.2 - 0.3*Age + 1.8*A_1(0) - 1.2*L(1) - 1.7*Y(0)$$

(m) Generate follow-up LNFAT using a normal distribution with data-specified mean

$(1.45 + 0.01*Age + 0.012*A_1(1) - 0.1*L(1) + 1.1*A_2(0) - 0.02*Y(0))$,

and variance $\sigma^2 = 0.04$

Summary of Changes to Certain Variables as Reflected in the Simulation Above

age -0.07-→ -0.3 and health -0.7 → -1.8 on LTPA1
age 0.006→0.03, and health –0.07→ -0.2 for LNFAT1

age -0.07 → -0.3, health -0.3→-1.2 and pf -0.7 → -1.7 for LTPA2
age 0.002→ 0.01, health 0.04→-0.1 , lnfat 0.46→1.1 for LNFAT2

Appendix 2a. SAS Code used for Simulation of Observed Data and the Unweighted and Weighted (IPTW) Regression. (Shown for Confounding of LNFAT and LTPA at Level of Spparcs Data).

```
/*****create a series of macros to generate variables for simulation**/

%macro age;
age=70 + sqrt(64)*rannor(433534);  *this is W;
if age<50 then do;
      age=50;
end;
age=round(age,1);
medage=age-70;    *median center age;
%mend age;

%macro hlth1(b0,b_age);
b0=&b0;
b_age=&b_age;
pred=1/(1+exp(b0 + b_age*medage));
health=0;
if pred > (ranuni(525252))  then do;
      health=1;
end;
drop b0 b_age;
%mend hlth1;

%macro ltpa1(b0,b_age,b_hlth);
b0=&b0;
b_age=&b_age;
b_hlth=&b_hlth;
ltpa=0;
pred2=1/(1+exp(b0 + b_age*medage + b_hlth*health));
if pred2 > (ranuni(525252))  then do;
      ltpa=1;
end;
drop b0 b_age b_hlth;
%mend ltpa1;

%macro lnfat1(b0,b_age,b_ltpa,b_hlth,sigma);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_hlth=&b_hlth;
sigmasq=&sigma;
lnfat=(b0 + b_age*medage + b_ltpa*ltpa +
b_hlth*health +  sqrt(SIGMASQ)*rannor(454523));
medlfat=lnfat - b0;          %*median center LNFAT;
drop b0 b_age b_ltpa b_hlth sigmasq;
%mend lnfat1;
```

```
%macro pf1(b0,b_age,b_ltpa,b_hlth,b_lnfat);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_hlth=&b_hlth;
b_lnfat=&b_lnfat;
pf=0;
pred4=1/(1+exp(b0 + b_age*medage + b_ltpa*LTPA  +
b_hlth*health + b_lnfat*medlfat));
if pred4>(ranuni(525252))  then do;
      pf=1;
end;
drop b0 b_age b_ltpa b_hlth b_lnfat;
%mend pf1;


%macro hlth2(b0,b_age,b_ltpa,b_lnfat,b_pf);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_lnfat=&b_lnfat;
b_pf=&b_pf;
 health2=0;
pred5=1/(1+exp(b0 + b_lnfat*medlfat + b_ltpa*ltpa +
b_pf*pf + b_age*medage));
if pred5 > (ranuni(525252))  then do;
      health2=1;
end;
drop b0 b_age b_ltpa b_lnfat b_pf;
%mend hlth2;

%macro ltpa2(b0,b_age,b_ltpa,b_pf,b_hlth);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_pf=&b_pf;
b_hlth=&b_hlth;
ltpa2=0;
pred6=1/(1+exp(b0 + b_age*medage + b_hlth*health2 +
b_ltpa*ltpa + b_pf*pf));
if pred6 > (ranuni(525252))  then do;
      ltpa2=1;
end;
drop b0 b_age b_ltpa b_pf b_hlth;
%mend ltpa2;

%macro lnfat2(b0,b_age,b_ltpa,b_hlth,b_lnfat,b_pf,sigma);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_hlth=&b_hlth;
b_lnfat=&b_lnfat;
b_pf=&b_pf;
sigmasq=&sigma;
lnfat2=(b0 + b_age*medage + b_ltpa*ltpa2 +
b_hlth*health2 + b_lnfat*medlfat + b_pf*pf + sqrt(sigmasq)*rannor(454523));
```

```
medlfat2=lnfat2 - b0;         %*median center LNFAT;
drop b0 b_age b_ltpa b_hlth b_lnfat b_pf sigmasq;
%mend lnfat2;

%macro pf2(b0,b_age,b_ltpa,b_hlth,b_lnfat,b_pf);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_hlth=&b_hlth;
b_lnfat=&b_lnfat;
b_pf=&b_pf;
pf2=0;
pred7=1/(1 + exp(b0 + b_age*medage + b_ltpa*LTPA2  +
b_hlth*health2 + b_lnfat*medlfat2 + b_pf*pf));
if pred7>(ranuni(525252))  then do;
      pf2=1;
end;
drop b0 b_age b_ltpa b_hlth b_lnfat b_pf;
       %mend pf2;
```

```
proc format;
value age3cat  low-64 = '<64'
              65-74 =  '65-74'
              74-high = '74-High';
run;
```

```
/*create a dataset with 1000 observations and Idn as subject identifier*/

data x0;
do idn=1 to 1000;
output;
end;
run;

/*** GENERATE TIME 1 VARIABLES**/

data x;
set x0;

%age

****GENERATE HEALTH(1);

%hlth1(0.2,-0.05)

****GENERATE LTPA(1);

%ltpa1(-0.5,0.07,0.7)
```

```
****GENERATE LNFAT(1);

%lnfat1(1.5,0.006,0.11,-0.07,0.15);

****GENERATE Y(1);

%pf1(1,-0.08,1.2,-0.4,0.8);

run;

/****check the distributions of variables from time1***/

proc univariate data=x plot;
var age lnfat;
run;

proc sort data=x;
by ltpa;
run;

proc univariate data=x;
by ltpa;
var age lnfat;
run;

proc sort data=x;
by health;
run;

proc univariate data=x;
by health;
var age lnfat;
run;

proc sort data=x;
by age;
run;

proc univariate data=x;
by age;
var lnfat;
format age age3cat.;
run;

proc freq data=x;
tables health ltpa pf health*ltpa health*pf ltpa*pf age*ltpa age*health
age*pf;
format age age3cat.;
run;

/****NOW GENERATE TIME 2 VARIABLES**/

data y;
set x;
```

```
*****GENERATE HEALTH(2);

%hlth2(-0.2,-0.05,0.8,0.2,-0.2);

*****GENERATE LTPA(2);

%ltpa2(1.2,0.07,-1.8,0.7,0.3)

*****GENERATE LNFAT(2);

%lnfat2(1.45,0.002,0.012,0.04,0.46,-0.02,0.04)

****GENERATE Y(2);

%pf2(1.7,-0.02,0.5,-0.5,0.4,-2.3);

run;

/*****check the distributions of variables from time 2***/

proc univariate data=y plot;
var  lnfat2;
run;

proc sort data=y;
by ltpa;
run;

proc univariate data=y;
by ltpa2;
var age lnfat2;
run;

proc sort data=y;
by health;
run;

proc univariate data=y;
by health2;
var age lnfat lnfat2;
run;

proc sort data=y;
by age;
run;

proc univariate data=y;
by age;
var lnfat2;
format age age3cat.;
run;

proc freq data=y;
tables health2 ltpa2 pf2 health2*ltpa2 health2*pf2 ltpa2*pf2 age*ltpa2
age*health2 age*pf2
health*health2 ltpa*ltpa2 pf*pf2 ltpa*health2 pf*health2;
```

```
format age age3cat.;
run;

data z;
set y(drop=pred pred2  pred4 pred5 pred6  pred7);
run;

proc sort data=z;
by idn;
run;

proc print data=z(obs=20);
title1 “20 obs from z”;
run;

****separate the single-line data into 2 lines per subject;

data a;
set z;
time=0;   output;
time=1; output;
run;

*******each line per subject will contain same variable
names---but indexed by time variable;

data a0;
set a(drop=health2 lnfat2 ltpa2 medlfat2 pf2);
if time=0;
run;

proc sort data=a0 (keep=idn lnfat ltpa medlfat pf) out=a00;
by idn;
run;

data a1;
set a(drop=health lnfat ltpa medlfat pf);
if time=1;
run;

data b;
set a0 a1(rename=(health2=health lnfat2=lnfat ltpa2=ltpa medlfat2=medlfat
pf2=pf));
run;

*****take some variables from 1st line for each subject and
merge it with the 2-line data so that past levels of LTPA
Leanfat and PF are available for modelling treatment at time2
---set to missing for 1st line;

proc sort data=b;
by idn;
run;
```

```
data b;
merge b(in=a) a00(in=b rename=(lnfat=pastfat ltpa=pastltpa
medlfat=pastmlft pf=pastpf));
by idn;
if a;
run;

data b;
set b;
if time=0 then do;
      pastfat=.;
      pastltpa=.;
      pastmlft=.;
      pastpf=.;
end;

run;

proc print data=b(obs=20);
title1 “20obs from working dataset”;
run;

****treatment models----check coeff against data-
generating model;

%macro a;

 data %do j=1 %to 2; b%eval(&j-1) %end;;
set b;
%do j=1 %to 2;
if time=%eval(&j-1) then output b%eval(&j-1);
%end;
run;

      %do j=1 %to 2;

            ods output  covparms=di%eval(&j-1);

            proc mixed data=b%eval(&j-1) method=ml;
            model lnfat=medage ltpa health
            %if %eval(&j>1)
            %then %do;
            pastmlft pastpf %end;/s outp=d%eval(&j-1);
            id idn;
            run;

      proc print data=d%eval(&j-1) (obs=10);
       title1 "check the data out of d%eval(&j-1)";
      run;

      proc print data=di%eval(&j-1);
       title1 "check the data out of di%eval(&j-1)";
      run;
```

```
            ods output  covparms=ni%eval(&j-1);

            proc mixed data=b%eval(&j-1) method=ml;
            model lnfat=
            %if %eval(&j>1)
            %then %do;
            pastmlft %end;/s outp=n%eval(&j-1);
            id idn;
            run;

      proc print data=n%eval(&j-1) (obs=10);
       title1 "check the data out of n%eval(&j-1)";
      run;

      proc print data=ni%eval(&j-1);
       title1 "check the data out of ni%eval(&j-1)";
      run;
proc sort data= d%eval(&j-1) (keep= idn Resid) out= d%eval(&j-1);
by idn;
run;

proc sort data= n%eval(&j-1) (keep= idn Resid) out= n%eval(&j-1);
by idn;
run;

proc sort data= b%eval(&j-1);
by idn;
run;

      data b%eval(&j-1);
      merge b%eval(&j-1)(in=a)
      d%eval(&j-1)(in=b rename=(Resid=rd))
      n%eval(&j-1)(in=b rename=(Resid=rn))
      ;
      by idn;
      if a and b;
      run;

      data b%eval(&j-1);
      set b%eval(&j-1);
      partnum=1;        %*little trick used in merging many obs w/ 1;
      run;              %*in the case of est variance from models;


      data di%eval(&j-1);
      set di%eval(&j-1);
      partnum=1;
      run;

      data ni%eval(&j-1);
      set ni%eval(&j-1);
      partnum=1;
      run;
```

```
      data b%eval(&j-1);
      merge b%eval(&j-1)(in=a) di%eval(&j-1)(in=b rename=(estimate=dest))
      ni%eval(&j-1)(in=b rename=(estimate=nest));
      by partnum;
      run;

 proc print data=b%eval(&j-1)(obs=20);
title1 “20obs from b%eval(&j-1)”;
run;

    proc sort data=b%eval(&j-1);
      by descending ltpa;
      run;

      proc logistic data=b%eval(&j-1) order=data;
      model ltpa=medage health
      %if %eval(&j>1) %then %do;
            pastltpa pastpf %end;/link=logit ;

         output out=md%eval(&j-1) predicted=metwd;
      run;

proc sort data= md%eval(&j-1);
by idn;
run;

      proc print data=md%eval(&j-1) (obs=10);
       title1 "check the data out of md%eval(&j-1)";
      run;

proc sort data= md%eval(&j-1) (keep= idn metwd) out= md%eval(&j-1);
by idn;
run;

      proc logistic data=b%eval(&j-1) order=data;
      model ltpa=
      %if %eval(&j>1) %then %do;
            pastltpa   %end;/link=logit ;
output out=mn%eval(&j-1) predicted=metwn;
      run;

proc sort data= mn%eval(&j-1);
by idn;
run;

proc print data=mn%eval(&j-1) (obs=10);
title1 "check the data out of mn%eval(&j-1)";
run;

proc sort data= mn%eval(&j-1) (keep= idn metwn) out= mn%eval(&j-1);
by idn;
run;

proc sort data= b%eval(&j-1);
by idn;
run;
```

```
      data b%eval(&j-1);
      merge b%eval(&j-1)(in=a)
      md%eval(&j-1)(in=b)
      mn%eval(&j-1)(in=c)
      ;
      by idn;
      if a and b;
      run;

      %end;


*****create the weights;

      data c;
      set b0 b1;

      lfat_dd=(1/sqrt(2*3.142*dest))*exp(-(rd**2)/(2*dest));
      lfat_nn=(1/sqrt(2*3.142*nest))*exp(-(rn**2)/(2*nest));
      swl=lfat_nn/lfat_dd;
      if ltpa=1 then do;
      met_sw=metwn/metwd;
      end;
      else if ltpa=0 then do;
      met_sw=(1-metwn)/(1-metwd);   *since pr(ltpa=1) being modelled;
                                  *pred prob of obs (ltpa=0) is 1-metwd;
      end;
      run;

proc print data=c(obs=20);
title1 "check the data with simple IPTW weights";
run;

proc univariate data=c;
var swl met_sw;
title1 "check distn of weights";
run;

data d1;
set c;
if time=0;
sw_0=swl;
met_0=met_sw;
run;

proc sort data=d1 (keep= idn sw_0 met_0) out=d1a;
by idn;
run;

proc sort data=c;
by idn;
run;
```

```
data d;
merge c(in=a) d1a(in=b);
by idn;
if a;
run;

proc print data=d(obs=20);
title1 “20obs with merge of time0 weights”;
run;

data e;
set d;
if time=1 then do;
      swl=swl*sw_0;
      met_sw=met_sw*met_0;
end;
wts=swl*met_sw;
 if wts gt 5 then do;
      wts=5;
 end;
 if wts lt 0.2 then do;
      wts=0.2;
 end;
run;

proc print data=e(obs=20);
title1 “20obs where weights calculated”;
run;

proc sort data=e;
by time;
run;

proc univariate data=e;
by time;
var swl met_sw wts;
title1 "check on wts by time";
run;

%mend a;
%a

proc genmod data=e descending;
class idn;
model pf=medlfat ltpa/dist=bin link=logit ;
repeated subject=idn/type=ind covb;
scwgt wts;
run;

proc genmod data=e  descending;
class idn;
model pf=medlfat ltpa/dist=bin link=logit ;
repeated subject=idn/type=exch covb;
title1 "without weights";
run;
```

```
proc genmod data=e  descending;
class idn;
model pf=medage health medlfat ltpa/dist=bin link=logit ;
repeated subject=idn/type=exch covb;
title1 "without weights adj for age and health";
run;
```

Appendix 2b. SAS Code used for Simulation of Bounded Full Data and Estimating "True" Parameters for the given MSM.

```
%macro age(ran);
age=70 + sqrt(64)*rannor(&ran);  *this is W;
if age<50 then do;
      age=50;
end;
age=round(age,1);
medage=age-70;    *median center age;
%mend age;

%macro hlth1(b0,b_age,ran);
b0=&b0;
b_age=&b_age;
pred=1/(1+exp(b0 + b_age*medage));
health=0;
if pred > (ranuni(&ran))  then do;
      health=1;
end;
drop b0 b_age;
%mend hlth1;

 %*****LNFAT and LTPA are given ----see below;

%macro pf1(b0,b_age,b_ltpa,b_hlth,b_lnfat,ran);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_hlth=&b_hlth;
b_lnfat=&b_lnfat;
pf=0;
pred4=1/(1+exp(b0 + b_age*medage + b_ltpa*LTPA  +
b_hlth*health + b_lnfat*medlfat));
if pred4>(ranuni(&ran))  then do;
      pf=1;
end;
drop b0 b_age b_ltpa b_hlth b_lnfat;
%mend pf1;

%macro hlth2(b0,b_age,b_ltpa,b_lnfat,b_pf,ran);
b0=&b0;
b_age=&b_age;
b_ltpa=&b_ltpa;
b_lnfat=&b_lnfat;
b_pf=&b_pf;
```

```
 health2=0;
pred5=1/(1+exp(b0 + b_lnfat*medlfat + b_ltpa*ltpa +
b_pf*pf + b_age*medage));
if pred5 > (ranuni(&ran))  then do;
      health2=1;
end;
drop b0 b_age b_ltpa b_lnfat b_pf;
%mend hlth2;

%macro pf2(b0,b_age,b_ltpa2,b_hlth,b_lnfat2,b_pf,ran);
b0=&b0;
b_age=&b_age;

b_ltpa2=&b_ltpa2;
b_lnfat2=&b_lnfat2;
b_hlth=&b_hlth;

b_pf=&b_pf;
pf2=0;
pred7=1/(1 + exp(b0 + b_age*medage + b_ltpa2*LTPA2  +
b_hlth*health2 + b_lnfat2*medlfat2 + b_pf*pf));
if pred7>(ranuni(&ran))  then do;
      pf2=1;
end;
drop b0 b_age  b_hlth  b_pf b_lnfat2 b_ltpa2;
%mend pf2;

%macro fullmac;

%*****10 randomly generated numbers from pocket calculator;

%let b1=7721;
%let b2=9897;
%let b3=2490;
%let b4=8668;
%let b5=343;
%let b6=5882;
%let b7=247;
%let b8=8067;
%let b9=7236;
%let b10=1033;

%do a=1 %to 10;

/*one can approximate the results of the same simulation*/
/*using 11 levels of LNFAT-see comment*/
/*beginning with the statement – "alternatively"*/

                        /*alternatively do k=1 to 11; do m=1 to 11*/

data a&a;
do i=1 to 1000; do j=1 to 2; do k=1 to 21; do l=1 to 2; do m=1 to 21;
output; end; end; end; end; end;
run;
```

```
proc print data=a&a(obs=100);
title1 "check 100obs from a&a";
title2 "creates 1764 records per subject---100 subjects";
run;

data b&a;
set a&a;
ltpa=j-1; ltpa2=l-1;
k=k-1;      m=m-1;

                /*alternatively lnfat=(k/5)+0.5;
                                lnfat2=(m/5)+0.5;*/

lnfat=(k/10)+0.5; lnfat2=(m/10)+0.5;
medlfat=lnfat-1.5; medlfat2=lnfat2-1.5;
if abs(medlfat-medlfat2)<=0.8;   %*bounds data—omits for every subject;
%if &a=1 %then %do;              %*differences in lnfat over t >0.8;
idn=i;
%end;
%else %do;
idn=%eval(&a*1000-1000)+i;
%end;
run;

proc datasets library=work nolist;
delete a&a;
run;

proc print data=b&a(obs=100);
title1 "check 100obs from b&a";
title2 "creates all counterfactuals per subject";
run;

data c&a;
do i=1 to 1000;
output;
end;
run;

data c&a;
set c&a;
%if &a=1 %then %do;
idn=i;
%end;
%else %do;
idn=%eval(&a*1000-1000)+i;
%end;
%age(&&b&a)
%hlth1(0.2,-0.05,&&b&a)
run;

proc print data=c&a(obs=100);
title1 "check 100obs from c&a";
title2 "generates 1 age and 1 health status";
title3 "for each subject";
run;
```

```
proc univariate data=c&a;
var age;
title1 "check distn of age in c&a data";
title2 "single line data";
run;

proc freq data=c&a;
tables health;
title1 "check distn of health in c&a data";
title2 "single line data";
run;

proc sort data=b&a;
by idn;
run;

proc sort data=c&a;
by idn;
run;

data d&a;
merge b&a(in=a drop=i j k l m) c&a(in=b keep=idn age medage health);
by idn;
if a and b;
run;

proc print data=d&a(obs=100);
title1 "check 100obs from d&a";
title2 "merges multiple b&a and single c&a";
run;

proc datasets library=work nolist;
delete b&a c&a;
run;

data e&a;
set d&a;
%pf1(1,-0.08,1.2,-0.4,0.8,&&b&a)
%hlth2(-0.2,-0.05,0.8,0.2,-0.2,&&b&a)
%pf2(1.7,-0.02,0.5,-0.5,0.4,-2.3,&&b&a)
run;

proc print data=e&a(obs=100);
title1 "check 100obs from e&a";
title2 "pf variable created";
run;

proc datasets library=work nolist;
delete d&a;
run;

proc print data=e&a(obs=50);
title1 "check 100obs from e&a";
title2 "before creating univariate formatted data";
run;
```

```
/*probably best to create permanent datasets here*/

data f&a(keep=disable lnfat vigor time idn);
set e&a;

%*create 2 lines for subject for modelling with GEE;

time=0; disable=pf; lnfat=medlfat; vigor=ltpa; output;
time=1; disable=pf2; lnfat=medlfat2; vigor=ltpa2; output;
run;

proc print data=f&a(obs=100);
title1 "check 100obs from f&a";
title2 "f&a";
run;

proc datasets library=work nolist;
delete e&a;
run;

%end;
quit;

%mend fullmac;
%fullmac

%macro bccs;

data sum(rename=(lnfat=medlfat vigor=ltpa disable=pf));
set %do i=1 %to 10; f&i %end;;

run;
proc genmod data=sum  descending;
class idn;
model pf=medlfat ltpa/dist=bin link=logit ;
repeated subject=idn/type=ind;
run;

 %mend bccs;
%bccs
```

Acknowledgement

This work was supported by Grant R01 AG09389 from the National Institute of Aging.

References

Neugebauer, R., van der Laan, M.J., 2002. Why Prefer Double Robust Estimates? Illustration with Causal Point Treatment Studies. U.C. Berkeley Division of Biostatistics Working Paper Series. Working paper 115.
http://www.bepress.com/ucbbiostat/paper115

Neugebauer, R., van der Laan, M.J., 2003. Locally Efficient Estimation of Nonparametric Causal Effects on Mean Outcomes in Longitudinal Studies. U.C. Berkeley Division of Biostatistics Working Paper Series. Working paper 134.
http://www.bepress.com/ucbbiostat/paper134

Robins, J.M., 1999. Association, Causation, and Marginal Structural Models. Synthese 121, 151-179.

Robins, J.M., Hernan, M.A., Brumback, B., 2000. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology 11(5), 550-560.

van der Laan, M.J., Robins, J.M., 2002. Unified Methods for Censored Longitudinal Data and Causality. Springer Verlag, New York.

Yu, Z., van der Laan, M.J., 2002a. Construction of Counterfactuals and the G-computation Formula. U.C. Berkeley Division of Biostatistics Working Paper Series. Working paper 122.
http://www.bepress.com/ucbbiostat/paper122

Yu, Z., van der Laan, M.J., 2002b. Double Robust Estimation in Longitudinal Marginal Structural Models. U.C. Berkeley Division of Biostatistics Working Paper Series. Working paper 132.
http://www.bepress.com/ucbbiostat/paper132