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

Year 2007 *Paper* 214

The Causal Effect of Recent Leisure-Time Physical Activity on All-Cause Mortality Among the Elderly

Oliver Bembom* Mark J. van der Laan†

Ira B. Tager‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, bembom@gmail.com

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

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

http://biostats.bepress.com/ucbbiostat/paper214

The Causal Effect of Recent Leisure-Time Physical Activity on All-Cause Mortality Among the Elderly

Oliver Bembom, Mark J. van der Laan, and Ira B. Tager

Abstract

We analyze data collected as part of a prospective cohort study of elderly people living in and around Sonoma, CA, in order to estimate, for each round of interviews, the causal effect of leisure-time physical activity (LTPA) over the past year on the risk of mortality in the following two years. For each round of interviews, this effect is estimated separately for subpopulations defined based on past exercise habits, age, and whether subjects have had cardiac events in the past. This decomposition of the original longitudinal data structure into a series of point-treatment data structures corresponds to an application of history-adjusted marginal structural models as introduced by van der Laan et al. (2005). We propose five different estimators of the parameter of interest, based on various combinations of the usual G-computation, inverse-weighting, and double robust approaches for the two layers of missingness corresponding to the treatment mechanism and right-censoring by drop-out. The models for all nuisance parameters required by these different estimators are selected data-adaptively. For most subpopulations, our analyses suggest that high leisure-time physical activity reduces the subsequent two-year mortality risk by about 50%. Among populations of elderly people aged 75 years or older, these effect estimates are generally significant at the 0.05 level. Notably, our analyses also identify one subpopulation that is estimated to experience an increase in mortality risk when exercising at a higher level, namely subjects aged 75 years or older with previous cardiac events and no history of habitual exercise (RR: 2.33, 95% CI: 0.76-4.35).

Contents

1 Introduction

A substantial body of epidemiological research indicates that recent and current physical activity in the elderly are associated with reductions in cardiovascular morbidity and mortality and improvement in or prevention of metabolic abnormalities that place elderly people at risk for these outcomes (CDC, 1989; van Dam et al., 2002; Lee et al., 2003; Esposito et al., 2003; Rosano et al., 2005). Data from studies of exercise physiology indicate that older, so-called master athletes retain a high level of fitness (Thomas et al., 1985). However, there are few epidemiological studies in general populations that have examined the relative contributions of habitual past and current physical activity on future morbidity and mortality. This issue is of relevance since there is increasing data that unhealthy lifestyles (lack of exercise, obesity, etc.) do increase the risk of metabolic abnormalities that are pro-atherogenic (Hu et al., 2001; Jacobs and Pereira, 2004; Kraus et al., 2002).

Tager et al. (1998) followed a group of people aged 55 years and older living in and around Sonoma, CA, over a time period of about ten years as part of a community-based longitudinal study of physical activity and fitness (Study of Physical Performance and Age Related Changes in Sonomans - SPPARCS). Our goal in analyzing the data that were collected as part of this study is to estimate the causal effect of recent leisure-time physical activity (LTPA) on all-cause mortality in an elderly population. We are particularly interested in estimating how this effect is modified by past exercise habits, age, and whether subjects have had cardiac events in the past. This would allow us to recommend levels of leisure-time physical activity that are tailored to these characteristics of a given person that put them at high risk for all-cause mortality.

2 Data structure

The SPPARCS study is a prospective cohort study of 2,092 people aged 55 years and older living in and around Sonoma, CA. Subjects were enrolled in 1993-1994 and followed through 2004. Each of the study participants was interviewed up to four times, with interviews spaced roughly two and a half years apart. Let M_0 denote the time at which the baseline interview was conducted. For $1 \leq t \leq 3$, let M_t give the time at which follow-up interview t was conducted.

At each interview, a questionnaire was used to assess leisure-time physical activity during the year preceding the interview. At the same time, a number of covariates were measured that might confound the relationship between leisure-time physical activity and mortality. These potential confounders include a lifetime profile of participation in leisure-time physical activity, a composite physical functioning score, self-rated health, smoking status, depression, BMI, living arrangement, as well as the presence or absence of a number of medical conditions (see tables 8 and 9 in the appendix for more details). Furthermore, subjects were asked whether they had previously experienced any of a number of cardiac events, a covariate that will be of interest to us as a potential effect modifier.

We define our treatment variable of interest $A_1(t)$ as an indicator for leisure-time physical

activity of 22.5 METs per week or greater during the year preceding interview t. This level of activity represents the minimum desired level of activity to maintain health, corresponding to brisk walking for 30 minutes at least five times a week. Since leisure-time physical activity is measured over the entire year preceding a given interview, it is possible that some of the potential confounders measured at the same interview have in fact been influenced by the subjects physical activity level over the past year. In particular, self-rated health, smoking status, and BMI are measured right at the interview; physical functioning is measured over the 1-month period preceding the interview; depression is measured over the 1-week period preceding the interview; and the presence of chronic health conditions is measured over the entire time period since the last interview. Hence, we cannot make the usual temporal ordering assumption that the covariates measured at a given time point precede the treatment variable at that time point and thus are not affected by that treatment variable. To be able to estimate causal effects, however, we have to define a data structure that respects such a temporal ordering assumption. We do so by defining $L(t)$ as the collection of covariates measured at interview $t-1$ rather than at interview t.

We define the collection of covariates preceding treatment at baseline, $L(0)$, as follows. At baseline, subjects were also asked to compare their current self-rated health to their self-rated health one year ago (better, same, worse). We use this information in conjunction with the subject's self-rated health at baseline to impute self-rated rated health one year prior to the baseline interview, which can now be included in $L(0)$. Furthermore, current smokers were asked when they began smoking, and ex-smokers were asked when they quit smoking. We use this information in conjunction with smoking status at baseline to impute smoking status one year prior to baseline to obtain another member of $L(0)$. Based on available questionnaire information, we are likewise able to impute the presence of lifetime cardiac events as well as a number of non-cardiac chronic health conditions one year prior to baseline. In this way, we can construct a collection of baseline covariates $L(0)$ that is known to precede baseline treatment $A_1(0)$ containing all covariates included in $L(t)$ for later time points except for the following: depression, physical functioning, living arrangement, and BMI.

Our effect modifiers of interest are defined as follows:

$$\begin{aligned} V_1(t) &= I(\text{Habitual exercise before study baseline}) \\ V_2(t) &= I(\text{cardiac event prior to interview } t-1) \\ V_3(t) &= I(\text{Age one year prior to interview } t \geq 75) \end{aligned}$$

While the variables V_1, V_2, and V_3 are available for all study participants, a number of the remaining variables in $L(t)$ are recorded incompletely. The missing values of such a variable $L_j(t)$ are imputed as follows: If past measurements for $L_j(t)$ are available, use the most recent past measurement to impute $L_j(t)$. Otherwise, if future measurements of $L_j(t)$ are available, use the closest future measurement to impute $L_j(t)$. Otherwise, use a typical value of $L_j(t)$ over the entire dataset to impute $L_j(t)$. For continuous $L_j(t)$, we use the median value of $L_j(t)$; for categorical $L_j(t)$, we use the mode of $L_j(t)$. $L_j(t)$ is then re-defined as $L_j(t) \equiv (L_j(t), \Delta_{L_j(t)})$, where $\Delta_{L_j(t))}$ is the indicator that $L_j(t)$ has been imputed rather

than measured. Tables 10 and 11 in the appendix summarize the results of this imputation procedure.

In this study, the treatment variable $A_1(t)$ may be missing at a given time point for one of two reasons: Either a subject has refused to participate in a given round of interviews, or a subject has not given information about his or her recent LTPA in spite of participating in an interview. The latter case happens most frequently when subjects choose to take the mail survey or a phone interview rather than the full home visit evaluation. We denote by D the time of the earliest interview with a missing value for $A_1(t)$. We enforce censoring of the treatment process to be monotone by discarding any data that might have been collected at times $t > D$ and denote the censoring process by $A_2(t) = I(D < M_t)$. Enforcing monotone censoring is necessary to avoid a practical violation of the ETA assumption since subjects who have refused to participate in a given round of interviews are very unlikely to return for the next round (there are only five such cases in the whole dataset). As a consequence, we are led to discard 87 out of 6,298 (1.4%) recorded treatment measurements that occurred after a previous missing treatment measurement.

At a given time point t, data become available in the following order. Based on the definition of $L(t)$ as the covariates measured at interview $t-1$, $L(t)$ will be available for any subject that has not dropped out or died by interview $t-1$. For any subjects who have not died by interview t, we next observe $A_2(t)$, i.e. whether or not they have dropped out at the current interview t. If this is not the case, we then record their treatment $A_1(t)$. Let $\tilde{T}_1 \equiv T \wedge C$ denote the time of the first occurrence of death and the end of the study period. Vital status data are available up to time $\tilde{T}_1$ for all subjects, even if they have dropped out prior to $\tilde{T}_1$. Let $\tilde{T}_2 \equiv T \wedge C \wedge D$ denote the time of the first occurrence of death, end of study, and drop-out. The treatment and covariate processes are then right-censored by $\tilde{T}_2$ and observed only through $\tilde{T}_2-$, the time point just prior to $\tilde{T}_2$. The observed data thus consist of n i.i.d. copies of

$$\begin{aligned} O &= (\tilde{T}_1 = T \wedge C,\ \Delta = I(T < C),\ \tilde{T}_2 = T \wedge C \wedge D,\ \bar{L}(\tilde{T}_2-),\ \bar{A}(\tilde{T}_2-)) \\ &= (\tilde{T}_1,\ \Delta,\ \tilde{T}_2,\ L(0),\ A_2(0),\ A_1(0),\ ...,\ L(\tilde{T}_2-),\ A_2(\tilde{T}_2-), A_1(\tilde{T}_2-)), \end{aligned}$$

where $A(t) \equiv (A_2(t), A_1(t))$.

3 Assumptions

Within the counterfactual framework for causal inference, we think of this observed data structure as a coarsened version of a full-data structure X that we would ideally have liked to observe. This full-data structure X consists of the collection of counterfactual covariate processes $\bar{X}_{\bar{a}_1} = (T_{\bar{a}_1},\ \bar{L}_{\bar{a}_1}(T_{\bar{a}_1}))$ with $\bar{a}_1$ ranging over the set $\mathcal{A}_1$ of possible treatment regimens. The observed data are now derived from the full data by two sequential coarsening steps. The first step is given by $Y = \varphi_1(X,\ A_1) = (\bar{A}_1,\ \bar{X}_{\bar{A}_1})$, i.e. Y contains only that particular covariate process $\bar{X}_{\bar{a}_1}$ corresponding to the actually observed treatment $\bar{A}_1$ (consistency assumption). The second step consists of censoring of this covariate process by drop-out D and end of study C: $O = \varphi_2(Y,\ C,\ D) = (\tilde{T}_1,\ \Delta,\ \tilde{T}_2,\ \bar{A}_1(\tilde{T}_2-)),\ \bar{X}_{\bar{A}_1}(\tilde{T}_2-))$.

We rely on the following additional standard assumptions that are necessary for causal effects to be identifiable from the observed data. First, we make the temporal ordering assumption $L_{\bar{a}_1}(t) = L_{\bar{a}_1(t-1)}(t)$ that states that covariates measured at time t are only affected by treatments at earlier time points. Our definition of $L(t)$ ensures that this assumption is met.

Next, we rely on the Sequential Randomization Assumption (SRA) for both the treatment mechanism and the drop-out mechanism. For the treatment mechanism, this assumption states that the choice $A_1(t)$ of treatment at time t is only affected by past treatment $\bar{A}_1(t-1)$ and measured covariates $\bar{L}(t)$:

$$g_1(a_1(t)|X,\ \bar{A}_1(t-1)) \equiv Pr(A_1(t) = a_1(t)|X,\ \bar{A}_1(t-1)) = g_1(a_1|\bar{A}_1(t-1),\ \bar{L}(t))$$

This corresponds to assuming that there are no unmeasured confounders of the relationship between treatment and mortality. Given the collection of potential confounders that we have measured, we are comfortable that this assumption is satisfied. We note that this assumption appears weaker at $t = 0$ since $L(0)$ does not contain all the potential confounders available at later time points. The drop-out mechanism likewise satisfies the SRA if the decision $A_2(t)$ to drop out at time t is only a function of past treatment $\bar{A}_1(t-1)$, past drop-out $\bar{A}_2(t-1)$, and measured covariates $\bar{L}(t)$:

$$g_2(a_2(t)|X,\ \bar{A}(t-1)) \equiv Pr(A_2(t) = a_2(t)|X,\ \bar{A}(t-1)) = g_2(a_2(t)|\bar{A}(t-1),\ \bar{L}(t))$$

Lastly, we make the Experimental Treatment Assignment (ETA) Assumption that states that there are essentially no values of the covariate process for which treatment or drop-out are assigned in a deterministic fashion:

$$Pr(A_1(t) = a_1|\ \bar{A}_1(t-1),\ \bar{L}(t)) > 0 \quad F_X - a.e. \text{ for } a_1 \in \{0, 1\}$$

and

$$Pr(A_2(t) = 0|\ \bar{A}(t-1),\ \bar{L}(t),\ A_2(t-1) = 0) > 0 \quad P_{F_x, g_1} - a.e.$$

This assumption is necessary for causal effects to be non-parametrically identifiable. Inverse-weighting based estimators rely crucially on this assumption since they do not posit a model for the full-data likelihood. Likelihood-based estimators rely entirely on extrapolation based on the posited model if this assumption is violated.

4 Parameter of interest

We are interested in estimating the causal effect of high leisure-time physical activity over the past year compared to lower activity on the risk of mortality over an appropriate time period, say the following two years. As mentioned above, we are particularly interested in estimating how this effect is modified by past exercise habits, age, and whether subjects have had cardiac events in the past. Since we have collected longitudinal data on our study participants, we might want to compare counterfactual mortality risks corresponding to treatment regimens

that prescribe high or low physical activity during the year preceding each of the interviews, with participants allowed to follow their observed activity patterns between interviews. The time at which follow-up interviews were conducted, however, varies greatly among the study participants, with the period between two interviews ranging from 274 to 1,410 days. In particular, this means that some study participants have had follow-up interviews within 2 years of the last interview whereas others have not. This makes counterfactual outcomes indexed by longitudinal treatment regimens as those described above difficult to interpret, especially since the timing of interviews might be related to a subject's characteristics. We therefore decide to consider counterfactual mortality outcomes that are only indexed by one treatment decision, essentially converting the longitudinal data structure into a point-treatment data structure. Thus we might be interested in estimating the parameter

$$\theta(a, v) = Pr(T_a \leq t_0 \mid V(0) = v),$$

where t_0 is the desired time frame of mortality, say 2 years, for all values of a and v.

For the sake of precision, however, we would still like to make use of the data that were collected at the follow-up interviews. This can be accomplished by considering corresponding parameters for later interviews and thus letting θ be indexed by the time j of the interview:

$$\theta(a, v, j) = Pr(T_{\bar{A}(j-1)a} \leq M_j + t_0 \mid T_{\bar{A}(j-1)} \geq M_j,\ V(j) = v)$$

Here $T_{\bar{A}(j-1)a}$ denotes the survival time for a subject whose LTPA regimen through interview $j-1$ corresponds to his or her observed activity pattern $\bar{A}(j-1)$ and whose physical activity at interview j was set to level a. For $j > 0$, this parameter gives the v-specific risk of mortality for the hypothetical scenario under which all study participants are allowed to follow a natural treatment regimen prior to interview j before then being assigned, for one year, to the treatment level of interest, a. Note that this counterfactual risk of mortality is conditional on having survived to interview j, but not on not having dropped out of the study before then.

We might now posit a model according to which $\theta(a, v, j)$ does not vary as a function of j, but corresponds to a common counterfactual mortality risk $\beta(a, v)$:

$$\theta(a, v, j) = \beta(a, v), j = 1, \ldots, 4 \tag{1}$$

We note, however, that the parameter $\beta(a, v)$ is not defined if this assumption fails. We therefore follow the approach developed in Neugebauer and van der Laan (2005b) to view 1 only as a working model that is used to define a smooth version of $\theta(a, v, j)$. This allows us to keep the model for the data-generating distribution non-parametric. Specifically, we define the parameter of interest $\beta(a, v)$ as a particular weighted average of the j-specific counterfactual mortality risk:

$$\begin{aligned} \beta(a, v) &= \frac{1}{\sum_{j=1}^{4} n(j, v)} \sum_{j=1}^{4} n(j, v)\theta(a, v, j) \\ &= \arg\min_{\beta(a,v)} \sum_{j=1}^{4} n(j, v)\Big[\theta(a, v, j) - \beta(a, v)\Big]^2 \end{aligned}$$

where $n(j,v) = \sum_{i=1}^{n} I(T_{\bar{A}(j-1),i} \geq M_{ji}, V_i(j) = v)$. The second equation shows that $\beta(a,v)$ can be viewed as a projection of the j-specific counterfactual mortality risks onto the space of functions of a and v alone, corresponding to the model 1, with weights given by $n(j,v)$. If model 1 holds, then $\beta(a,v)$ is equal to the common counterfactual mortality risk $\theta(a,v,j), j = 1,\ldots,4$. This assumption may not be so unrealistic since counterfactual mortality risks are stratified by age so that the aging of our cohort is, at least to some extent, taken into account. In the absence of this assumption, $\beta(a,v)$ becomes slightly harder to interpret, but at least remains a perfectly well-defined parameter.

We note that the approach we take here for pooling information across different time points can be viewed as an application of history-adjusted marginal structural models as introduced by van der Laan et al. (2005). These models have been proposed primarily for the purpose of studying time-dependent effect modification in longitudinal studies with a temporally meaningful baseline. van der Laan et al., for example, consider a cohort of AIDS patients that are enrolled in the study as soon as they lose virologic suppression. This allows one to place a new AIDS patients who has lost virologic suppression some known time t ago with respect to the time scale of the study and to recommend treatment decisions that have been derived specifically for patients whose virus has been non-suppressed for this amount of time. In the current context, however, the study baseline has no intrinsic temporal meaning since study participants are enrolled at no particular point in their lives. This precludes us from placing new subjects from the same study population with respect to the time scale of the study and to base recommendations specifically on that temporal placement. Instead, we apply the same models and estimators as those proposed by van der Laan et al. for the sake of gaining precision by pooling information across several time points.

5 Estimators

We propose the following five estimators for the parameter $\beta(a,v)$.

5.1 G-computation

In the absence of censoring, the G-computation estimator for $\theta(a,v,j)$ would be given by

$$\theta_n(a,v,j) = P_n\Big\{Q_{1n}(a,j)\Big|T_{\bar{A}(j-1)} \geq M_j, V(j) = v\Big\}$$

where we use the notation

$$P_n f(O) \equiv \sum_{i=1}^{n} f(O_i)$$

for the empirical mean, and $Q_{1n}(a,j) = \hat{Pr}(T_{\bar{A}(j-1)a} \leq M_j + t_0 | T_{\bar{A}(j-1)} \geq M_j,\ \bar{L}(j),\ \bar{A}_1(j))$ is an estimate of the conditional risk of mortality in the time period following interview j given the observed treatment and covariate processes through interview j among subjects who were alive at interview j.

In the presence of right-censoring by drop-out, $Q_1(a,j)$ is still identifiable from the data since drop-out is assumed to be independent of the full data given the observed past so that

$$Pr(T_{\bar{A}(j-1)a} \leq M_j + t_0 | T_{\bar{A}(j-1)} \geq M_j,\ \bar{L}(j),\ \bar{A}_1(j)) =$$
$$Pr(T_{\bar{A}(j-1)a} \leq M_j + t_0 | T_{\bar{A}(j-1)} \geq M_j,\ \bar{L}(j),\ \bar{A}_1(j), A_2(j) = 0)$$

Hence we can simply estimate $Q_1(a,j)$ among those subjects who still remain in the study at interview j. The empirical mean of $Q_{1n}(a,j)$, however, can only be evaluated over those same subjects, a subset of the target population that is unlikely to be representative of the whole population, making the estimator infeasible. This problem can be addressed by simulating realizations of the observed data structure through time point j for those subjects who have dropped out prior by time point j. The empirical mean of $Q_{1n}(a,j)$ can then be taken over this partially imputed data set O^b to obtain an estimate $\theta_n^b(a,v,j)$ of $\theta(a,v,j)$. Repeating this procedure a sufficient number of times and averaging over b leads to the G-computation estimator

$$\theta_n^{G-comp}(a,v,j) = \sum_{b=1}^{B} \theta_n^b(a,v,j)$$

An estimate of $\beta(a,v)$ can then obtained as

$$\beta_n^{G-comp}(a,v) = \frac{1}{\sum_{j=1}^{4} n(j,v)} \sum_{j=1}^{4} n(j,v)\theta_n^{G-comp}(a,v,j). \tag{2}$$

Note that only the treatment and covariate processes need to be simulated since the failure process is observed through time $\tilde{T}_1$ for all study participants. Realization of the treatment process can be generated based on an estimate g_{1n} of the treatment mechanism. To generate realizations of the covariate process, we need to estimate the distributions $Q_2(j) = Pr(L(j) \mid \bar{L}(j-1),\ \bar{A}(j-1))$, $j = 1, \ldots, 4$. We here use a dimension reduction approach based on propensity scores (Rosenbaum and Rubin, 1983) to simplify this task. Let

$$e_1(t) = e_1(\bar{L}(t-1),\ \bar{A}(t-1),\ A_2(t)) \equiv Pr(A_1(t) = 1 |\ \bar{L}(t-1),\ \bar{A}(t-1),\ A_2(t))$$

$$e_2(t) = e_2(\bar{L}(t-1),\ \bar{A}(t-1)) \equiv Pr(A_2(t) = 1 |\ \bar{L}(t-1),\ \bar{A}(t-1))$$

be the propensity scores corresponding to the treatment mechanism and the drop-out mechanism, respectively. If the original data structure satisfies the assumption of no unmeasured confounders, then the same is true for the lower-dimensional data structure

$$R = (\tilde{T}_1,\ \Delta,\ \tilde{T}_2,\ e_1(0),\ e_2(0),\ A(0),\ ...,\ e_1(\tilde{T}_2-),\ e_2(\tilde{T}_2-),\ A(\tilde{T}_2-))$$

in which the covariate process $L(t)$ consist of only two components at each time point. We can thus estimate the nuisance parameters $Q_2(j) = Pr(L(j) |\ \bar{L}(j-1),\ \bar{A}(j-1))$ for this lower-dimensional data structure. Since the dimension reduction step relies on estimation of g_2, the estimate $\beta_n^{G-comp}(a,v)$ is consistent only if all of the nuisance parameters Q_1, Q_2, g_1, and g_2 are estimated consistently.

5.2 G-comp-IPCW

An alternative approach to salvaging the simple point treatment estimator is based on inverse-probability-of-censoring weights that give small weights to subjects who are unlikely to have been censored by interview j and large weights to subjects who have not dropped out by interview j, but would have been likely to do so, given their history. This approach leads to the estimator

$$\theta_n^{G-comp-IPCW}(a,v,j) = P_n\Big\{\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_{2n}(\bar{A}_2(j)=\mathbf{0}|X)}Q_{1n}(a,j)\Big|T_{\bar{A}(j-1)} \geq M_j, V(j)=v\Big\},$$

This estimator corresponds to the estimating function

$$D(a,v,j \mid \theta(a,v,j)) = I(T_{\bar{A}(j-1)} \geq M_j, V(j)=v)\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_2(\bar{A}_2(j)=\mathbf{0}|X)}Q_1(a,j) - \theta(a,v,j),$$

which is unbiased since

$$E\Big[I(T_{\bar{A}(j-1)} \geq M_j, V(j)=v)\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_2(\bar{A}_2(j)=\mathbf{0}|X)}Q_1(a,j)\Big] - \theta(a,v,j) =$$

$$E\Big[\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_2(\bar{A}_2(j)=\mathbf{0}|X)}Q_1(a,j)\Big|T_{\bar{A}(j-1)} \geq M_j, V(j)=v\Big] - \theta(a,v,j) =$$

$$E\Big[E\Big[\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_2(\bar{A}_2(j)=\mathbf{0}|X)}Q_1(a,j)\Big|X\Big]\Big|T_{\bar{A}(j-1)} \geq M_j, V(j)=v\Big] - \theta(a,v,j) =$$

$$E\Big[\frac{Q_1(a,j)}{g_2(\bar{A}_2(j)=\mathbf{0}|X)}E\Big[I(\bar{A}_2(j)=\mathbf{0})\Big|X\Big]\Big|T_{\bar{A}(j-1)} \geq M_j, V(j)=v\Big] - \theta(a,v,j) =$$

$$E\Big[\frac{Q_1(a,j)}{g_2(\bar{A}_2(j)=\mathbf{0}|X)}g_2(\bar{A}_2(j)=\mathbf{0}|X)\Big|T_{\bar{A}(j-1)} \geq M_j, V(j)=v\Big] - \theta(a,v,j) =$$

$$E\Big[Q_1(a,j)\Big|T_{\bar{A}(j-1)} \geq M_j, V(j)=v\Big] - \theta(a,v,j) = 0$$

An estimate of $\beta(a,v)$ can then obtained as

$$\beta_n^{G-comp-IPCW}(a,v) = \frac{1}{\sum_{j=1}^4 n(j,v)}\sum_{j=1}^4 n(j,v)\theta_n^{G-comp-IPCW}(a,v,j). \tag{3}$$

This estimate is consistent only if both the nuisance parameters Q_1 and g_2 are estimated consistently.

5.3 IPTW-IPCW estimator

An alternative to the two likelihood-based approaches described above consists of applying the general estimating-function methodology for semi-parametric missing-data models satisfying coarsening at random (CAR) described in van der Laan and Robins (2003). This approach is based on first finding estimating functions that could be used if the data were completely observed and then mapping them into estimating functions that can be used for the actually observed, coarsened data structure, as described in Theorem 1.3 of van der Laan and Robins. In the present context, we will have to carry out this mapping step twice, first to map estimating functions for the completely-observed data structure X into estimating functions for the data structure Y, and then to map those into estimating functions for the observed data structure O. At first glance, Theorem 1.3 of van der Laan and Robins appears not to apply to the initial mapping step since the counterfactuals considered here, allowing subjects to follow their observed treatment through time point $j-1$, make the parameter of interest a function of the treatment mechanism as well as the full data structure. The hypothesis of the theorem requiring the parameter of interest to be a function of the full data structure alone can still be seen to hold, however, by conceiving for each time point j of a full data structure X^j that is equivalent to a point-treatment full data structure in which the baseline covariates contain $\bar{L}(j)$ as well as $\bar{A}(j-1)$ (van der Laan et al., 2005).

An unbiased full-data estimating function for $\beta(a,v)$ is given by

$$D(X,\beta(a,v)) = \sum_{j=1}^{4} D_j(X,\beta(a,v)) = \sum_{j=1}^{4} I_1(j,v)\Big[I(T_{\bar{A}(j-1)a} \leq t_{0j}) - \beta(a,v)\Big],$$

where $I_1(j,v) = I(T_{\bar{A}(j-1)} \geq M_j, V(j) = v)$ and $t_{0j} = M_j + t_0$. Since our model for the data-generating distribution is non-parametric, the tangent space is locally saturated, the orthogonal complement of the full-data nuisance tangent space is given by

$$\Lambda^{Full,\perp} = \{aD(X,\beta(a,v) : a \in \mathbb{R}\},$$

and $D(X,\beta(a,v))$ is in fact an efficient estimating function. The simplest mapping from this full-data estimating function to an observed-data estimating function consists of two sequential applications of weights, inverse-probability-of-treatment weights (IPTW) followed by inverse-probability-of-censoring weights (IPCW). The resulting j-specific IPTW-IPCW estimating function is given by

$$D_j^{IPTW-IPCW}(O,\beta(a,v)|g_1,g_2) = D_j(X,\beta(a,v))\frac{I(A_1(j)=a)}{g_{1j}(a|X)}\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_2(\mathbf{0}|X)}.$$

The estimating equation

$$0 = \sum_{i=1}^{n} D_j^{IPTW,IPCW}(O_i,\beta_n^{IPTW-IPCW}(a,v)|g_1,g_2)$$

can be written as

$$0 = \sum_{i=1}^{n} f^{IPTW-IPCW}(O_i) - \beta_n^{IPTW-IPCW} \sum_{i=1}^{n} g^{IPTW-IPCW}(O_i),$$

where

$$\begin{aligned} f^{IPTW-IPCW}(O_i) &= \sum_{j=1}^{4} I_{1,i}(j,v) I(T_{\bar{A}(j-1)a,i} \leq t_{0j,i}) \frac{I(A_{1,i}(j)=a)}{g_{1j}(a|X_i)} \frac{I(\bar{A}_{2,i}(j)=\mathbf{0})}{g_2(\mathbf{0}|X_i)} \\ g^{IPTW-IPCW}(O_i) &= \sum_{j=1}^{4} I_{1,i}(j,v) \frac{I(A_{1,i}(j)=a)}{g_{1j}(a|X_i)} \frac{I(\bar{A}_{2,i}(j)=\mathbf{0})}{g_2(\mathbf{0}|X_i)}. \end{aligned}$$

Hence the estimator $\beta_n^{IPTW-IPCW}(a,v)$ is given by

$$\beta_n^{IPTW-IPCW}(a,v) = \frac{\sum_{i=1}^{n} f^{IPTW-IPCW}(O_i)}{\sum_{i=1}^{n} g^{IPTW-IPCW}(O_i)} \tag{4}$$

This estimator is consistent only if the two nuisance parameters g_1 and g_2 are estimated consistently.

5.4 DR-IPCW estimator

Alternatively, we may apply a double robust mapping for the first coarsening step, corresponding to the treatment mechanism, followed by an inverse-weighting mapping for the second coarsening step, corresponding to the drop-out mechanism. For this purpose, we first obtain the j-specific double-robust estimating function in the absence of censoring by projecting the corresponding IPTW estimating function

$$D_j^{IPTW}(Y, \beta(a,v)|g_1) = I_1(j,v)\Big\{\frac{I(A(j)=a)}{g_{1j}(a|X)} I(T_{\bar{A}(j-1)a} \leq t_{0j}) - \beta(a,v)\Big\},$$

onto the orthogonal complement of the nuisance tangent space Λ_1^j associated with the treatment mechanism at time j. Λ_1^j consists of all functions of $\bar{L}(j)$ and $\bar{A}(j)$ with conditional mean zero given $\bar{L}(j)$, $\bar{A}(j-1)$, and $A_2(j)$:

$$\Lambda_1^j = \{\varphi(\bar{L}(j), \bar{A}(j)) - E[\varphi(\bar{L}(j), \bar{A}(j)) \mid \bar{L}(j), \bar{A}(j-1), A_2(j)] : \varphi\}$$

The projection of D_j^{IPTW} onto the orthogonal complement of Λ_1^j can be obtained by subtracting from D_j^{IPTW} its projection onto Λ_1^j. This latter projection can be computed by first finding the conditional expectation of D_j^{IPTW} given $\bar{L}(j)$ and $\bar{A}(j)$ and then subtracting the expectation of that quantity over the conditional distribution of $A_1(j)$ given $\bar{L}(j)$, $\bar{A}(j-1)$, and $A_2(j)$. This leads to the following double robust estimating function in the absence of censoring:

$$D_j^{DR}(Y, \beta(a,v)|g_1, Q_1) = I_1(j,v)\Big\{\frac{I(A(j)=a)}{g_{1j}(a|X)}\Big[I(T_{\bar{A}(j-1)a} \leq t_{0j}) - Q_1(a,j)\Big] + Q_1(a,j) - \beta(a,v)\Big\},$$

where, as before, $Q_1(a,j) = Pr(T_{\bar{A}(j-1)a} \leq t_{0j} | T_{\bar{A}(j-1)} \geq M_j,\ \bar{L}(j),\ \bar{A}(j))$. Applying an inverse-weighting mapping to this estimating function, we obtain the j-specific DR-IPCW estimating function

$$D_j^{DR,IPCW}(O, \beta(a,v)|g_1, Q_1, g_2) = D_j^{DR}(Y, \beta(a,v)|g_1, Q_1)\frac{I(\bar{A}_2(j) = \mathbf{0})}{g_2(\bar{A}_2 = \mathbf{0}|Y)}$$

The estimating equation

$$0 = \sum_{i=1}^{n} D_j^{DR,IPCW}(O_i, \beta_n^{DR-IPCW}(a,v)|g_1, Q_1)$$

can be written as

$$0 = \sum_{i=1}^{n} f^{DR-IPCW}(O_i) - \beta_n^{DR-IPCW} \sum_{i=1}^{n} g^{DR-IPCW}(O_i),$$

where

$$f^{DR-IPCW}(O_i) =$$

$$\sum_{j=1}^{4} I_{1,i}(j,v)\Big[\frac{I(A_{1,i}(j) = a)}{g_{1j}(a|X_i)}[I(T_{\bar{A}(j-1)a,i} \leq t_{0j,i}) - Q_{1,i}(a,v)] + Q_{1,i}(a,v)\Big]\frac{I(\bar{A}_{2,i}(j) = \mathbf{0})}{g_2(\mathbf{0}|X_i)}$$

and

$$g^{DR-IPCW}(O_i) = \sum_{j=1}^{4} I_{1,i}(j,v)\frac{I(\bar{A}_{2,i}(j) = \mathbf{0})}{g_2(\mathbf{0}|X_i)}.$$

Hence the estimator $\beta_n^{DR-IPCW}(a,v)$ is given by

$$\beta_n^{DR-IPCW}(a,v) = \frac{\sum_{i=1}^{n} f^{DR-IPCW}(O_i)}{\sum_{i=1}^{n} g^{DR-IPCW}(O_i)}. \tag{5}$$

This estimator is consistent if g_2 as well as least one of g_1 and Q_1 are estimated consistently.

5.5 DR-DR estimator

Lastly, we may apply a double robust mapping to $D_j^{DR}(Y, \beta(a,v)|g_1, Q_1)$ in the hope of obtaining an estimating function that is orthogonal to both the treatment and the drop-out mechanism. This is accomplished by subtracting from $D_j^{DR-IPCW}$ its projection onto the nuisance tangent spaces Λ_2^l, $l = 1, \ldots, j$, that are associated with the drop-out mechanisms at all time points l up to j. Λ_2^l consists of all functions of $\bar{L}(l)$, $\bar{A}(l-1)$, and $A_2(l)$ with conditional mean zero given $\bar{L}(l)$ and $\bar{A}(l-1)$:

$$\Lambda_2^j = \{\varphi(\bar{L}(l), \bar{A}(l-1), A_2(l)) - E[\varphi(\bar{L}(l), \bar{A}(l-1), A_2(l)) \mid \bar{L}(l), \bar{A}(l-1)] : \varphi\}$$

The projection of $D_j^{DR-IPCW}$ onto Λ_1^l can be obtained by first finding the conditional expectation of $D_j^{DR-IPCW}$ given $\bar{L}(l)$, $\bar{A}(l-1)$, and $A_2(l)$ and then subtracting the expectation of that quantity over the conditional distribution of $A_2(l)$ given $\bar{L}(l)$ and $\bar{A}(l-1)$. This leads to the estimating function

$$\begin{aligned} D_j^{DR,DR}(O,\beta(a,v)|g_1,Q_1,g_2,Q_2) &= D_j^{DR,IPCW}(O,\beta(a,v)|g_1,Q_1,g_2) \\ &\quad -\sum_{l=1}^{j}\Big[Q_{31}(a,v,j,l)-\beta(a,v)Q_{32}(v,j,l)\Big]dM_2(j) \end{aligned}$$

where

$$\begin{aligned} Q_{32}(v,j,l) &= E\Big[I_1(j,v)\frac{I(\bar{A}_2(j)=\mathbf{0})}{g_2(\bar{A}_2=\mathbf{0}|X)}\Big|\bar{L}(l),\bar{A}(l),A_2(l)\Big], \\ Q_{31}(a,v,j,l) &= E\Big[D_j^{DR,IPCW}\Big|\bar{L}(l),\bar{A}(l),A_2(l)\Big]+\beta(a,v)Q_{32}(v,j,l), \\ dM_2(j) &= I(A_2(j)=0)-g_{2j}(0|Y). \end{aligned}$$

The expectations $Q_{31}(a,v,j,l)$ and $Q_{32}(v,j,l)$ must be estimated numerically by Monte-Carlo simulation that relies as before on a data reduction step based on propensity scores. In this case, however, we need to generate realizations of the entire observed data structure so that we also rely on an estimate of the full failure process, a nuisance parameter we denote by Q_4.
The estimating equation

$$0=\sum_{i=1}^{n}D_j^{DR,DR}(O_i,\beta_n^{DR-DR}(a,v)|g_1,Q_1,g_2,Q_2)$$

can be written as

$$0=\sum_{i=1}^{n}f^{DR-DR}(O_i)-\beta_n^{DR-DR}\sum_{i=1}^{n}g^{DR-DR}(O_i),$$

where

$$f^{DR-DR}(O_i)=$$

$$\sum_{j=1}^{4}\Big\{I_{1,i}(j,v)\Big[\frac{I(A_{1,i}(j)=a)}{g_{1j}(a|X_i)}[I(T_{\bar{A}(j-1)a,i}\le t_{0j,i})-Q_{1,i}(a,v)]+Q_{1,i}(a,v)\Big]\frac{I(\bar{A}_{2,i}(j)=\mathbf{0})}{g_2(\mathbf{0}|X_i)}$$

$$-\sum_{l=1}^{j}Q_{21,i}(a,v,j,l)dM_2(j)\Big\}$$

and

$$g^{DR-DR}(O_i)=\sum_{j=1}^{4}\Big\{I_{1,i}(j,v)\frac{I(\bar{A}_{2,i}(j)=\mathbf{0})}{g_2(\mathbf{0}|X_i)}-\sum_{l=1}^{j}Q_{22,i}(v,j,l)dM_2(j)\Big\}$$

Hence the estimator $\beta_n^{DR-DR}(a,v)$ is given by

$$\beta_n^{DR-DR}(a,v) = \frac{\sum_{i=1}^n f^{DR-DR}(O_i)}{\sum_{i=1}^n g^{DR-DR}(O_i)}. \tag{6}$$

This estimator is consistent if g_2 as well as least one of g_1 and Q_1 are estimated consistently. It is not double robust with respect to the drop-out mechanism since estimation of Q_2 itself relies on estimation of g_2.

5.6 Nuisance parameter models

All nuisance parameters are modelled data-adaptively and separately for each time point j. The treatment and drop-out mechanisms g_1 and g_2 consist of regressions of an indicator variable on $\bar{L}(j)$ and $\bar{A}_1(j-1)$. These two nuisance parameters are modelled data-adaptively using the `polyclass()` function of a model selection algorithm based on polynomial spline functions (Kooperberg et al., 1997). The nuisance parameter $Q_2(j)$ consists of regressions of continuous propensity scores on $\bar{L}(j-1)$ and $\bar{A}(j-1)$. These are estimated data-adaptively using the `polymars()` function based on the same algorithm.

The nuisance parameter $Q_1(j,a)$ consists of a regression of a failure indicator on $\bar{L}(j)$ and $\bar{A}_1(j)$ and could thus also be modelled using `polyclass()`. Models selected in this way, however, are likely to contain neither the treatment variable $A_1(j)$ nor any interaction terms between $A_1(j)$ and the effect modifiers of interest $V(j)$. Such models are unsatisfactory since they do not allow us to examine the impact of $A_1(j)$ on the risk of mortality and the dependence of this impact on $V(j)$. To explicitly acknowledge that we are interested in estimating the effect of $A_1(j)$ on the risk of mortality within strata of $V(j)$, we might hence fit separate data-adaptive regression models for each stratum of $V(j)$ and each value of A. This is problematic, however, for the following reason. Suppose $(\bar{L}(j), \bar{A}_1(j-1))$ contains an important confounder that is very strongly correlated with $A_1(j)$ and that has an independent effect on the risk of mortality. Then clearly this variable should be included in a model predicting mortality risk from $A_1(j)$ and $(\bar{L}(j), \bar{A}_1(j-1))$ to adequately control for confounding. Within groups defined by $A_1(j)$, this variable will show very little variation, however, and thus will contribute little to the accurate estimation of mortality risk. Model selection procedures are thus unlikely to include this variable in the chosen regression model. We therefore adopt the following two-step approach: First, we fit a data-adaptive regression model to estimate the risk of mortality as a function of $(\bar{L}(j), \bar{A}_1(j-1))$ alone, excluding $A_1(j)$ from the set of candidate explanatory variables. Then we manually add $A_1(j)$, the interaction terms $A_1(j) \times V(j)$, as well as any terms in $V(j)$ that have not yet been selected to the identified model. The first step guarantees that no important confounders are omitted due to strong correlations with $A_1(j)$. The second step then ensure that the model contains all terms of interest.

The nuisance parameter Q_4, representing the failure process, is estimated in an analogous fashion. We first obtain a data-adaptive estimate of the hazard of failure as a function of $(\bar{L}(t), \bar{A}_1(t-1))$ alone by using the `hare()` function of the same model selection algorithm

as above. We then add the same terms as for the model of Q_1 and fit the corresponding Cox proportional-hazards model to obtain the desired estimate of Q_4.

5.7 Comparison of the five estimators

Table 1 compares these five estimators with respect to their dependence on the various nuisance parameters.

Estimator	$g_1(a_1\|X)$	$Q_1(a,j)$	$g_2(a_2\|Y)$	$Q_2(j)$
G-comp	✓	✓	✓	✓
G-comp-IPCW		✓	✓	
IPTW-IPCW	✓		✓	
DR-IPCW	✓ $\longleftrightarrow$	✓	✓	
DR-DR	✓ $\longleftrightarrow$	✓	✓	

Table 1: Comparison of estimators in their dependence on nuisance parameters. An estimator relies on each nuisance parameter marked by ✓. A ✓$\longleftrightarrow$✓signifies that the estimator depends on consistent estimation of at least one of the two corresponding nuisance parameters.

We note that consistency of the G-computation estimator implies consistency of all four other estimators, i.e. these estimators are guaranteed to be consistent whenever the G-computation estimator is consistent. Likewise, consistency of either the G-comp-IPCW estimator or the IPTW-IPCW estimator implies consistency of the DR-IPCW and DR-DR estimators. These last two estimators are hence maximally robust with respect to misspecification of nuisance parameter models among the candidate estimators we consider here.

5.8 Inference

Inference for all five candidate estimators can be based on the bootstrap. For the three estimating-function based estimators, we may also turn to arguments based on influence curves. For the DR-DR estimator, the influence curve is straightforward to derive, assuming that all nuisance parameters are estimated consistently. In that case, the estimating function lies in the orthogonal complement of the nuisance tangent space so that the corresponding influence curve is given by an appropriately scaled version of the estimating function itself:

$$IC_0^{DR-DR}(O,\beta(a,v)|g_1,Q_1,g_2,Q_2) = c^{-1}D^{DR-DR}(O,\beta(a,v)|g_1,Q_1,g_2,Q_2)$$

where

$$c = -\frac{\partial}{\partial\beta(a,v)}E\Big[D(O,\beta(a,v))\Big]\Big|_{\beta(a,v)=\beta_0(a,v)} = E\Big[g^{DR-DR}(O)\Big].$$

Thus

$$\sqrt{n}(\beta_n^{DR-DR} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IC_0^{DR-DR}(O_i, \beta(a,v)|g_1, Q_1, g_2, Q_2) + o_P(1)$$

and in particular

$$\sqrt{n}(\beta_n^{DR-DR} - \beta_0) \stackrel{D}{\Longrightarrow} N(0, \Sigma)$$

where

$$\Sigma = Var(IC_0^{DR-DR}(O, \beta(a,v)|g_1, Q_1, g_2, Q_2)).$$

The influence curve can be estimated by plugging in estimates for the parameter of interest as well as the nuisance parameter. The scaling factor c can be estimated by the empirical mean of $g^{DR-DR}(O)$. This leads to an estimate of the variance-covariance matrix Σ so that inference can then be based on the asymptotic normal distribution displayed above. Inference obtained in this way relies however on the assumption that the nuisance parameters are estimated at a sufficiently fast rate. This assumption holds if nuisance parameters are estimated within parametric models, but becomes somewhat questionable when data-adaptive model selection techniques are employed. Hence we suspect that influence-curve based inference may be optimistic in this case.

The influence curve of the IPTW-IPCW and DR-IPCW estimators are not given by a simple scaling of the estimating function itself since that estimating function is not orthogonal to all nuisance parameters. The estimating function of the IPTW-IPCW estimator can be obtained by projecting the scaled version

$$IC_0^{IPTW-IPCW}(O, \beta(a,v)|g_1, g_2) = c^{-1} D^{IPTW-IPCW}(O, \beta(a,v)|g_1, g_2)$$

onto the orthogonal complement of the nuisance tangent spaces associated with the treatment and drop-out mechanism. The estimating function of the DR-IPCW estimator can likewise be obtained by projecting the scaled version

$$IC_0^{DR-IPCW}(O, \beta(a,v)|g_1, Q_1, g_2) = c^{-1} D^{DR-IPCW}(O, \beta(a,v)|g_1, Q_1, g_2)$$

onto orthogonal complement of the nuisance tangent space for the drop-out mechanism. Since projections are norm-reducing, the variance of the true influence curves is no smaller than the variance of these preliminary influence curves. Hence they can generally be used for conservative inference about the parameter of interest. In the present context of data-adaptively selected nuisance parameter models, however, it is unclear how this conservative behavior might balance with the optimistic behavior due to second-order terms. We investigate this question by comparing confidence intervals for the three estimating-function based estimators that are based on the bootstrap to those based on the influence curve arguments presented here.

6 Results

6.1 Study population

Table 2 summarizes the number of subjects remaining in the study at each time point as well as the number of subjects lost to death and drop-out since the previous interview. Table 3 characterizes the various populations of interest we obtain when we pool subjects across the four different time points. Participants under the age of 75 with no previous cardiac events make up more than half of the entire sample, but account for less than 20% of all observed deaths. Participants with previous cardiac events on the other hand represent slightly more than 15% of the sample, but account for more than 30% of deaths. On they whole, the sample is fairly well balanced between the two LTPA regimens under consideration, with slightly more participants in the high-activity group. Among young subjects, cardiac events appear more commonly in the male population than in the female population, with the gap becoming less pronounced in the older population. As to be expected, populations with previous cardiac events as well as older populations are characterized by a decrease in physical functioning scores as well as in self-perceived health as compared to younger and healthier populations.

Table 2: Subjects lost to death and drop-out as well as remaining population at each time point.

Interview	Deaths	Drop-outs	Population
1	0	0	2074
2	89	237	1748
3	139	253	1356
4	76	182	1098

Table 3: Summary statistics for populations of interest. Populations are defined based on age, the presence (hrt+) or absence (hrt-) of previous cardiac events, and habitual (hab+) or non-habitual (hab-) exercise patterns in the past.

	Age<75				**Age≥75**			
	hrt-,hab-	hrt-,hab+	hrt+,hab-	hrt+,hab+	hrt-,hab-	hrt-,hab+	hrt+,hab-	hrt+,hab+
Mortality								
Deaths	45	15	13	9	112	63	61	25
Population	2116	1288	312	182	1183	646	344	205
LTPA								
<22.5 METs	39.7%	24.5%	40.1%	33%	55.6%	42.9%	55.5%	39%
≥22.5 METs	60.3%	75.5%	59.9%	67%	44.4%	57.1%	44.5%	61%
Sex								
Female	65.4%	57.1%	42%	40.7%	65.1%	57.4%	54.7%	40.5%
Male	34.6%	42.9%	58%	59.3%	34.9%	42.6%	45.3%	59.5%
Age								
0th %ile	53	54	55	55	75	75	75	75
25th %ile	63	62	66	65	77	77	77	77
50th %ile	67	67	70	69	80	80	81	80
75th %ile	71	71	72	72	84	84	84	83
100th %ile	75	75	75	75	98	103	98	95
NRB								
0th %ile	0	0	0.17	0.16	0	0	0	0
25th %ile	0.84	0.84	0.74	0.77	0.6	0.65	0.53	0.62
50th %ile	1	1	0.9	1	0.84	0.88	0.78	0.84
75th %ile	1	1	1	1	1	1	1	1
100th %ile	1	1	1	1	1	1	1	1
Health								
Excellent health	38.5%	44.4%	17.6%	20.9%	29.3%	31.7%	10.5%	14.6%
Good health	49%	43.6%	55.8%	41.2%	53.2%	48.8%	52.6%	51.2%
Fair health	10.4%	10.2%	21.5%	31.9%	14.1%	14.6%	28.5%	23.9%
Poor health	2.1%	1.9%	5.1%	6%	3.4%	5%	8.4%	10.2%

6.2 Selected nuisance parameter models

Tables 12 through 22 in the appendix summarize the selected nuisance parameter models. The treatment model for the baseline interview consists mainly of measures of physical activity between age 40 and the beginning of the study along with age, sex, and an indicator for a decline in physical activity over the five to ten years preceding the study. Treatment models for subsequent interviews focus very heavily on LTPA measurements from previous interviews, incorporating in each case the entire available LTPA history. Other variables taken into account by these models are physical functioning scores, BMI, and age. Since past LTPA measurements as well physical functioning scores are not available at the first time point, but appear very prominently in the treatment models for later time points, we suspect that the SRA is more closely approximated for later time points in the study than for earlier ones.

The selected models for the drop-out mechanism are all fairly simple, with old age being the only recurring risk factor for drop-out. Since no other major predictors of mortality nor LTPA appear in these models, we suspect that adjusting for right-censoring by drop-out will only have a minor impact on our estimates.

The models selected for the two-year risk of mortality focus primarily on available physical functioning measurements, with decreased scores associated, as expected, with an increased mortality risk. Other variables selected include sex, an indicator for chronic disease, and an indicator for a decline in physical activity in the five to ten years preceding the study. Again we are led to suspect that the SRA is more closely approximated for later time points than for the baseline interview since physical functioning scores are not available for the first time point. Note that the absence of observed failures in some of the populations of interest at the third and fourth time points leads to large coefficient estimates with even larger standard errors for some of the interaction terms in the corresponding models. This is not a problem here since the estimated coefficients only represent nuisance parameters that are not immediately used to obtain estimates for the parameters of interest.

6.3 Mortality and relative risk estimates

Figures 1 and 2 as well as table 6 in the appendix show the estimated counterfactual mortality risks for the two different LTPA levels and the various subpopulations of interest. Figures 3 and 4 as well as table 7 in the appendix summarize these estimates in the form of relative risks of mortality for comparing high-level activity to low-level activity.

On the whole, the five estimators agree fairly well with each other, with two notable exceptions among subjects aged 75 years and older with previous cardiac events. Among such subjects with a history of habitual exercise, the three estimating-function based estimates for the high-activity LTPA regimen lie around 35% while the two likelihood-based estimates are closer to 15%. Among the same subjects with no history of habitual exercise, the three estimating-function based estimates for the low-activity LTPA regimen lie around 20% while the two likelihood-based estimates are closer to 13%.

As described above in section 5.7, the DR-IPCW and DR-DR estimators are more robust

than the two likelihood-based estimators in the sense that consistency of either of the latter two estimators implies consistency of the former two estimators, but not vice versa. Recall also that the DR-DR estimator, just like the DR-IPCW estimator, relies on a consistent estimate of the drop-out mechanism g_2 since estimation of the additional nuisance parameter Q_2 itself relies on estimation of g_2. The DR-DR estimator is hence no more robust than the DR-IPCW estimator. If Q_2 is estimated consistently, the DR-DR estimator can typically be expected to be more efficient than the DR-IPCW estimator. In the present case, however, we note that the DR-DR estimator appears to be more variable than the DR-IPCW estimator, suggesting that Q_2 is in fact not estimated consistently. This may not be too surprising considering that estimation of Q_2 is based on simulations that require a normality assumption for all conditional covariate distributions as well as correct models for the treatment and drop-out mechanisms that are needed for the propensity-score based dimension reduction. Since the DR-IPCW estimator hence appears to enjoy the most desirable robustness and efficiency properties among the estimators considered here, we will focus on results obtained based on this approach.

The general pattern of estimated two-year mortality risks agrees well with what one might expect on the basis subject-matter considerations. Thus, estimated mortality risk are considerably lower in the young population than in the old population. Within a given age group, participants with previous cardiac events are consistently estimated to be at higher mortality risks than those without previous cardiac events. Likewise, subpopulations with a history of habitual exercise are typically estimated to be at lower mortality risks than comparable subpopulations with no such history. Two-year mortality risk estimates range from 1% (95% CI: 0-3%) for young participants with a history of habitual exercise and no previous cardiac events following the high-activity LTPA regimen to 36% (95% CI: 13-67%) for old participants with no history of habitual exercise and previous cardiac events following the high-activity LTPA regimen.

Among young subjects, our analysis suggests that high LTPA reduces the two-year risk of mortality by about 40% (Table 7, rows 1-4). The relative risk estimates for young subjects who reported past habitual exercise are very imprecise, undoubtedly due to the small number of subjects and deaths in these groups (Table 3), with somewhat stronger statistical evidence for a beneficial effect of LTPA on mortality in the remaining two groups. Since the point estimates are similar across all four groups of young subjects, we are led to believe that this effect is not modified in a significant way by previous cardiac events or past habitual exercise.

Among older subjects with a history of previous cardiac events, the point estimates for the relative risk of mortality were lower and more precise than for the younger group (Table 7, rows 5-6). The findings for older subjects with previous cardiac events were somewhat surprising (Table 7, rows 7-8). For those with a positive history of habitual exercise, the relative risk of mortality for high-level LTPA compared to low-level LTPA was estimated as 0.35 (95% CI: 0.00-1.20). For those without a history of habitual exercise, the same relative risk was estimated as 2.33 (95% CI: 0.76-4.35%). Apart from the possibility of representing a chance increase in mortality risk, this finding could be due to the fact that high-level LTPA was undertaken too soon after the non-fatal cardiac event and could have thus contributed

to increased mortality in unconditioned individuals. We did not have data to address this speculation.

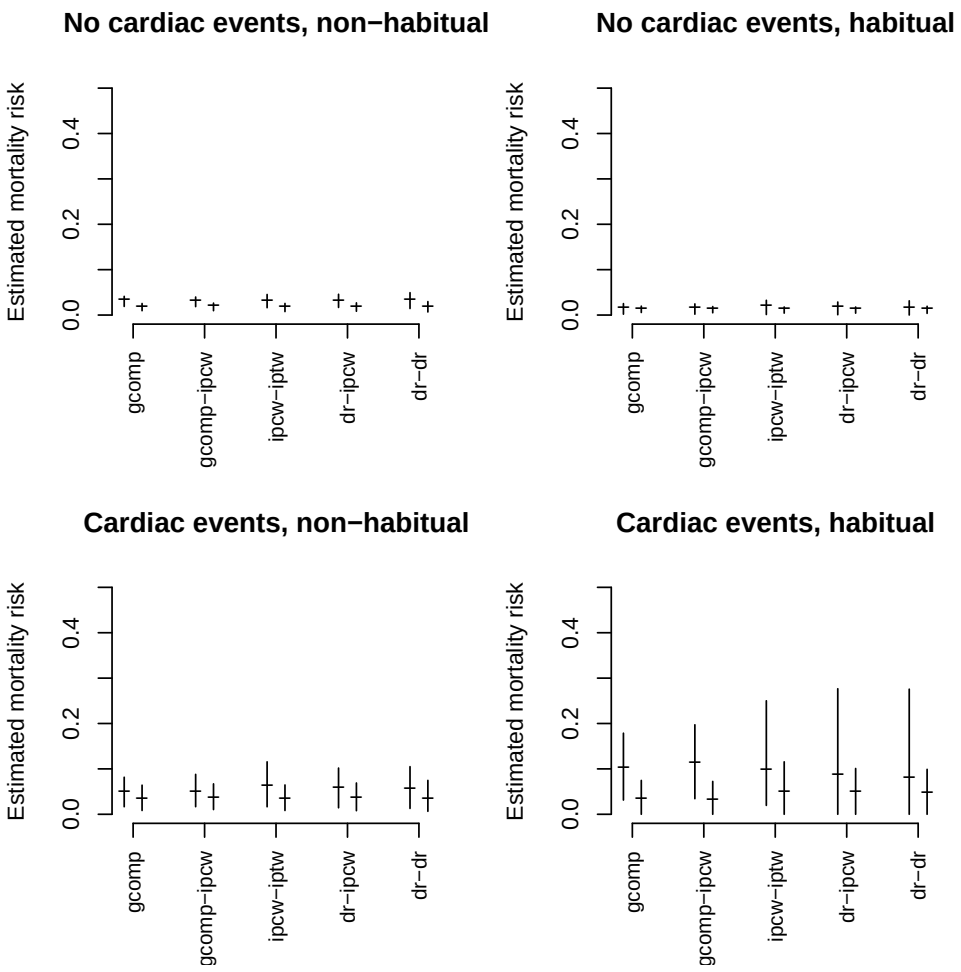

Figure 1: Mortality risk estimates with 95% bootstrap confidence intervals for age < 75. For each subpopulation and estimator, the plot shows estimated counterfactual mortality risks under low LTPA on the left and those under high LTPA on the right.

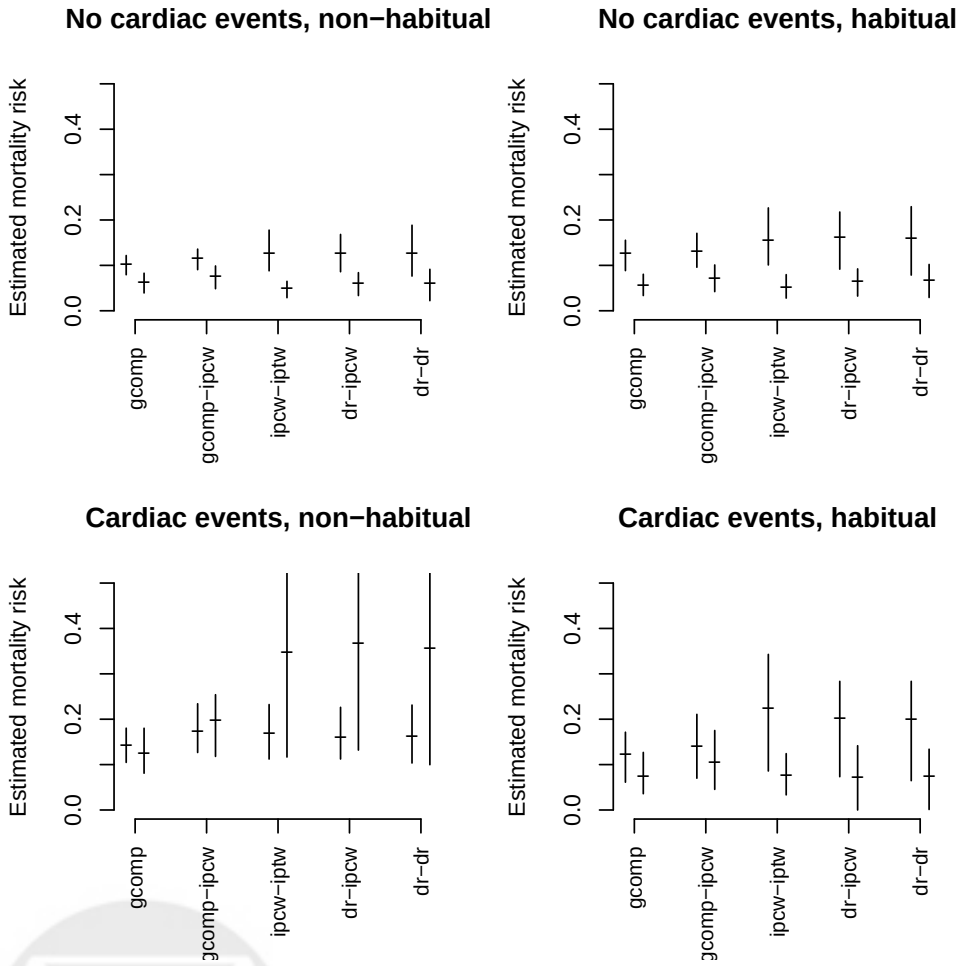

Figure 2: Mortality risk estimates with 95% bootstrap confidence intervals for age ≥ 75. For each subpopulation and estimator, the plot shows estimated counterfactual mortality risks under low LTPA on the left and those under high LTPA on the right.

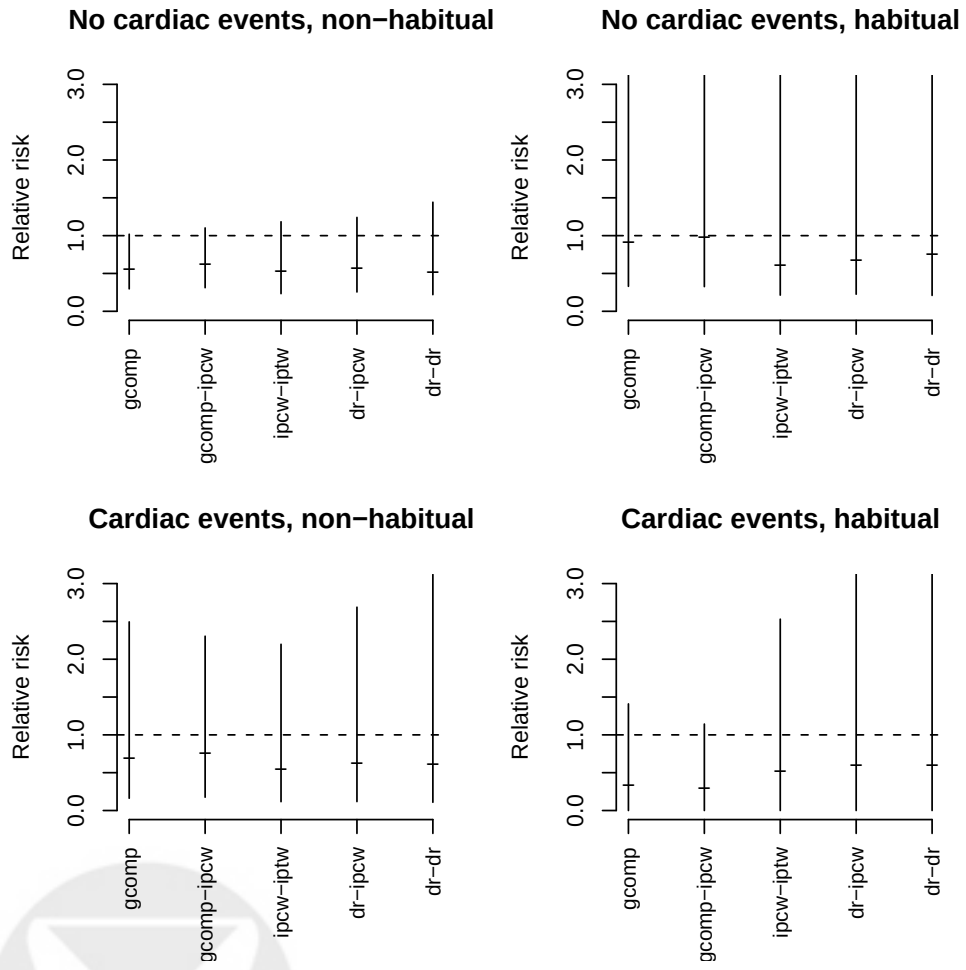

Figure 3: Relative risk estimates with 95% bootstrap confidence intervals for age < 75.

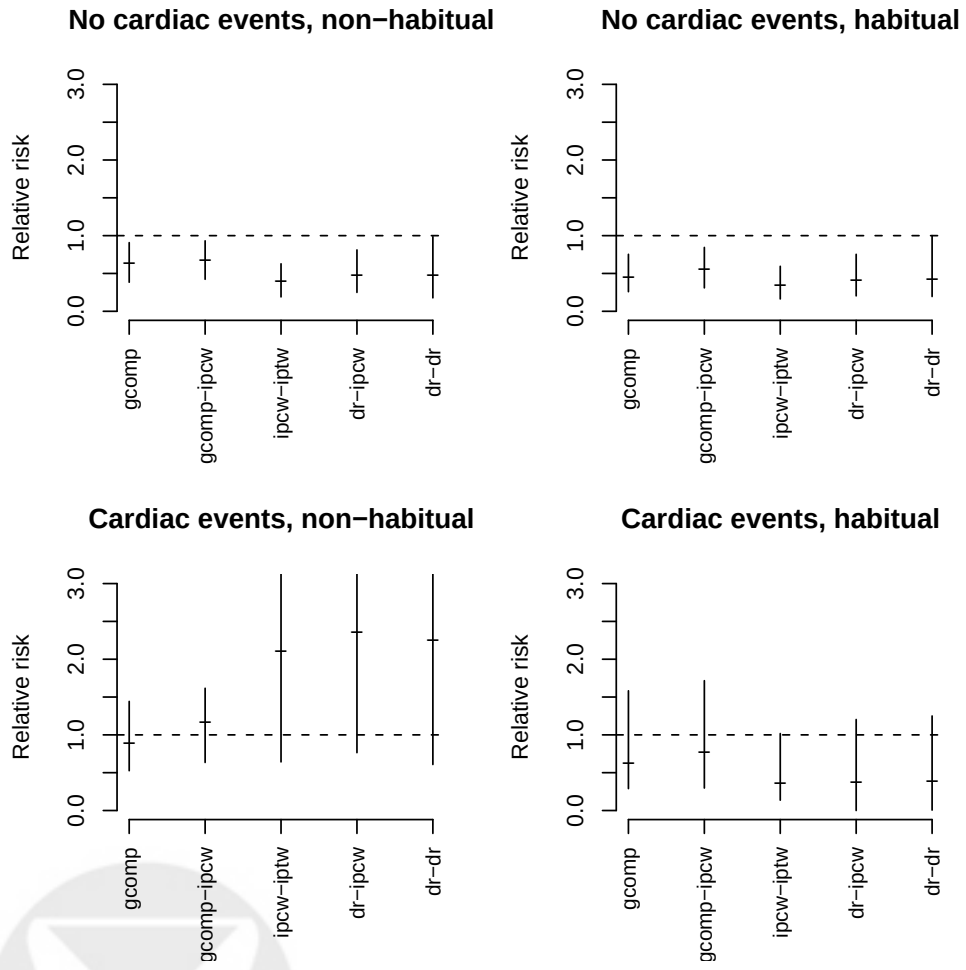

Figure 4: Relative risk estimates with 95% bootstrap confidence intervals for age ≥ 75.

6.4 Assessing the validity of the ETA assumption

As mentioned in section 3, the desired causal effects are only non-parametrically identifiable if the Experimental Treatment Assignment assumption holds. In the absence of this assumption, likelihood-based estimators rely entirely on extrapolation based on the posited model for the full-data likelihood, and inverse-weighting based estimators typically provide biased results. In fact, these latter estimators also perform poorly if the ETA assumption is practically violated, i.e. if for some values of the baseline covariates, treatment is assigned in a nearly deterministic fashion (Neugebauer and van der Laan, 2005a). In our analyses, the IPTW-IPCW, DR-IPCW, and DR-DR estimators require that the treatment mechanism satisfy the ETA assumption, while the G-computation and G-comp-IPCW estimators may still yield consistent estimates through extrapolation; the G-comp-IPCW, IPTW-IPCW, DR-IPCW, and DR-DR estimators all require that the drop-out mechanism satisfy the ETA assumption, with only the G-computation estimator enjoying the potential of yielding consistent estimates through extrapolation.

Wang et al. (2006) propose to use the following simulation-based approach to examine the extent to which inverse-weighting-based estimates might be biased due to such a violation of the ETA assumption: One first obtains estimates of both the full-data generating distribution as well as the treatment and censoring mechanisms. These estimates now allow one to simulate realizations of the observed data structure. For this data-generating distribution, the true values of the parameters of interest can be computed through G-computation. On the other hand, one can obtain a sampling distribution of inverse-weighting-based estimates by applying the inverse-weighting-based estimator to each simulated realization of the observed data structure. Since the SRA is trivially satisfied in this case, any discrepancy between the mean of these estimates and the true parameter value has to be due to a violation of the ETA assumption.

We employ this approach here to explore the extent to which the DR-IPCW estimator might be afflicted by bias due to such a violation of the ETA assumption. Figure 5 shows representative sampling distributions of the DR-IPCW estimates of the log-odds of mortality for some values of the stratification variables of interest. Table 4 summarizes the estimated bias of the DR-IPCW point estimates for the two counterfactual mortality risks as well as the relative risk of mortality. It also shows the relative risk estimate we obtain by correcting for the estimated bias along with similarly bias-corrected confidence intervals; these corrected confidence intervals ignore the variability due to estimating the bias and are hence only given as a rough indication of the variability in the bias-corrected point estimates. The bias for most estimates appears to be quite small, with somewhat larger estimated biases for the relative risk estimates than the mortality estimates. Considerable biases of the relative risk estimates are only observed for the population of younger subjects with a history of habitual exercise. Since the relative risk estimates for this group of subjects are highly variable, the 1,000 iterations used for the simulation study may be too small to obtain an accurate estimate of the mean the corresponding sampling distribution. We might speculate that the estimated bias would decrease as we increase the number of iterations, but this is largely a moot point since the high variability of these relative risk estimates already precludes us from placing

much trust in them. Lastly, we note that in the majority of cases, the estimated bias for the relative risk estimates is towards the null value of 1.0, with bias-corrected estimates further away from 1.0. This simulation study thus suggests that most of the previous findings are in fact supported by stronger evidence than is reflected in the initial confidence intervals.

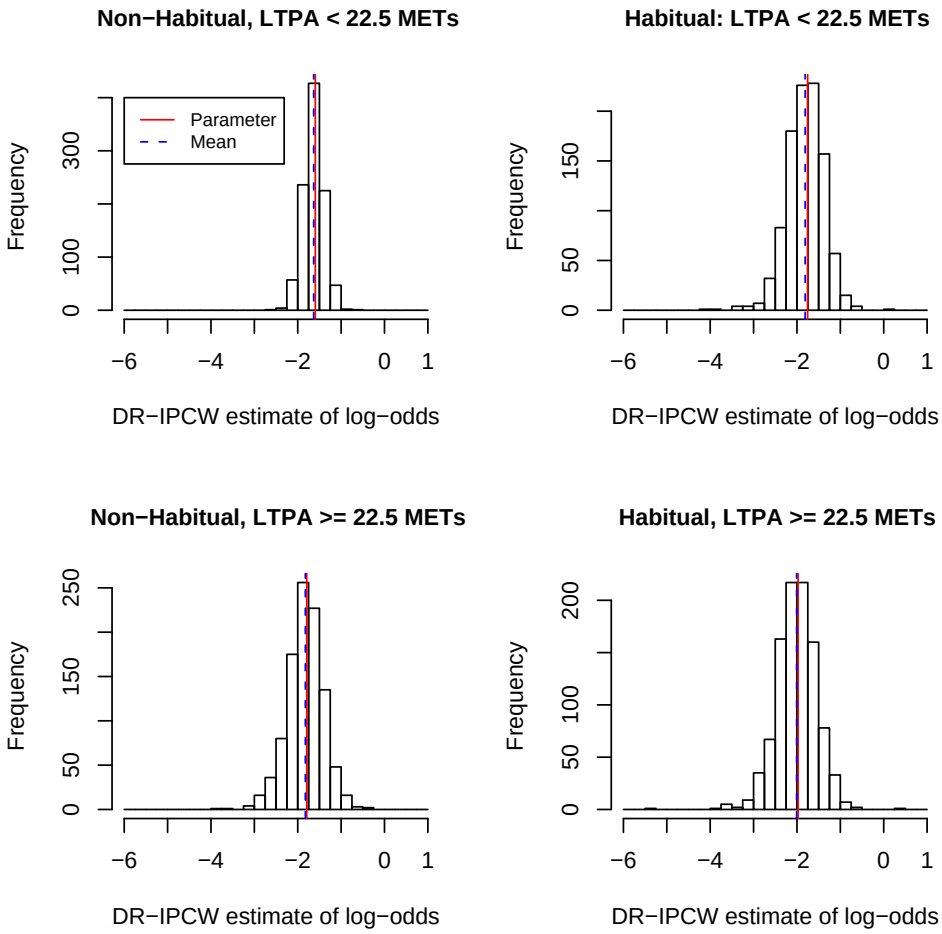

Figure 5: Sampling distribution of DR-IPCW estimates of the log-odds of mortality among participants aged 75 years or older with previous cardiac events and habitual or non-habitual exercise habits in the past under the two different LTPA regimens. The blue vertical line indicates the mean of a given distribution, and the red vertical line indicates the true parameter value for the simulation experiment.

Table 4: Estimated bias of DR-IPCW estimates (DR-IPCW point estimates, bias as percentage of point estimates) for both counterfactual mortality risks as well as the relative risk of mortality along with the bias-corrected relative risk estimate with similarly corrected 95% confidence interval.

old	card	hab	Low LTPA		High LTPA		RR		Corrected RR	
			0.0007	(0.0291, 2.41%)	-0.0003	(0.0159, -1.89%)	-0.019	(0.547, -3.5%)	0.567	(0.264, 1.286)
		✓	-0.0010	(0.0148, -6.76%)	0.0008	(0.0097, 8.25%)	0.200	(0.655, 30.5%)	0.502	(0.173, 9.636)
	✓		-0.0005	(0.0557, -0.90%)	0.0037	(0.0334, 11.08%)	0.015	(0.599, 2.5%)	0.585	(0.115, 2.623)
	✓	✓	0.0008	(0.0829, 0.97%)	0.0031	(0.0471, 6.58%)	0.572	(0.568, 100.6%)	0.283	(0.000, Inf)
✓			-0.0007	(0.1223, -0.57%)	0.0006	(0.0554, 1.08%)	-0.005	(0.453, -1.1%)	0.458	(0.252, 0.820)
✓		✓	-0.0058	(0.1577, -3.68%)	0.0028	(0.0608, 4.61%)	0.035	(0.386, 9.0%)	0.354	(0.187, 0.691)
✓	✓		-0.0027	(0.1556, -1.74%)	0.0033	(0.3630, 0.91%)	-0.007	(2.333, -0.3%)	2.341	(0.767, 4.362)
✓	✓	✓	0.0006	(0.1978, 0.30%)	0.0043	(0.0689, 6.24%)	0.007	(0.348, 1.9%)	0.342	(0.000, 1.181)

6.5 Mortality risk estimates stratified by interview

Recall that the parameter of interest $\beta(a, v)$ is defined as a particular weighted average of interview-specific counterfactual mortality risks $\theta(a, v, j)$. This pooling across time points increases our effective sample size and thus improves the precision of our estimates. As described in section 4, $\beta(a, v)$ is well-defined even if $\theta(a, v, j)$ is not constant as a function of j. The interpretation of $\beta(a, v)$ would be simplified considerably, however, if this assumption were to hold. Hence, we next look at the how the interview-specific mortality estimates change as a function of time.

Figures 6 and 7 show the DR-IPCW estimates of counterfactual mortality risks separately for each time point t. Considering the variability of these estimates, there appears to be little evidence that mortality risks change as a function of time. Since the cohort ages throughout the study period, one concern might be, for instance, that mortality risks rise steadily over time in spite of the crude adjustment for age in the form of an indicator variable for age greater than 75 years. The plots, however, do not reveal a consistent trend of this sort.

We note that among subjects at least 75 years of age with previous cardiac events but no history of habitual exercise, the counterfactual mortality risk for high LTPA is estimated to be considerably higher for the third time point than for all remaining time points. While the data for time points two and four also support an increase in risk for the high-activity regimen, with the first time point suggesting no change in risk, the magnitude of the estimated 2.33-fold increase in risk appears to be driven primarily by the estimates for the third time point. As noted previously, we have reason to believe that the SRA is more closely approximated for the later time points, with treatment and failure models able to capitalize on more abundant information regarding recent activity patterns and physical functioning levels. This observation suggests that not too much weight should be given to the estimates for the first time point, which stand alone in indicating no impact of increased LTPA on mortality risk. We are still interested, however, in understanding what might explain the unusually high mortality estimate for the high-activity regimen at the third time point. For this purpose, table 3 characterizes the populations of subjects aged 75 years and older with previous cardiac events and no history of habitual exercise for each of the four time points. The table illustrates the confounding of the relationship between LTPA and mortality by physical functioning levels, as NRB scores are sharply lower in the group of subjects exercising at less than 22.5 METs a week. For time points one, two, and four, the crude mortality estimate is considerably lower for the group of subjects exercising at a high level as compared to the group of subjects exercising at a lower level. The adjusted estimates in figure 7 indicate that, once we account for physical functioning scores and other confounders, this relationship is in fact reversed. We now note that even the crude analysis finds a considerably higher risk for the high-activity regimen at the third time point than at any of the other time points, with the risk even exceeding that of the low-activity regimen for the same time point. This suggests that the unusually high corresponding adjusted mortality estimate is unlikely to be due to problems with our adjustment for confounding, but rather to the small numbers of subjects encountered in this population for each given time point and the resulting instability in the crude estimates.

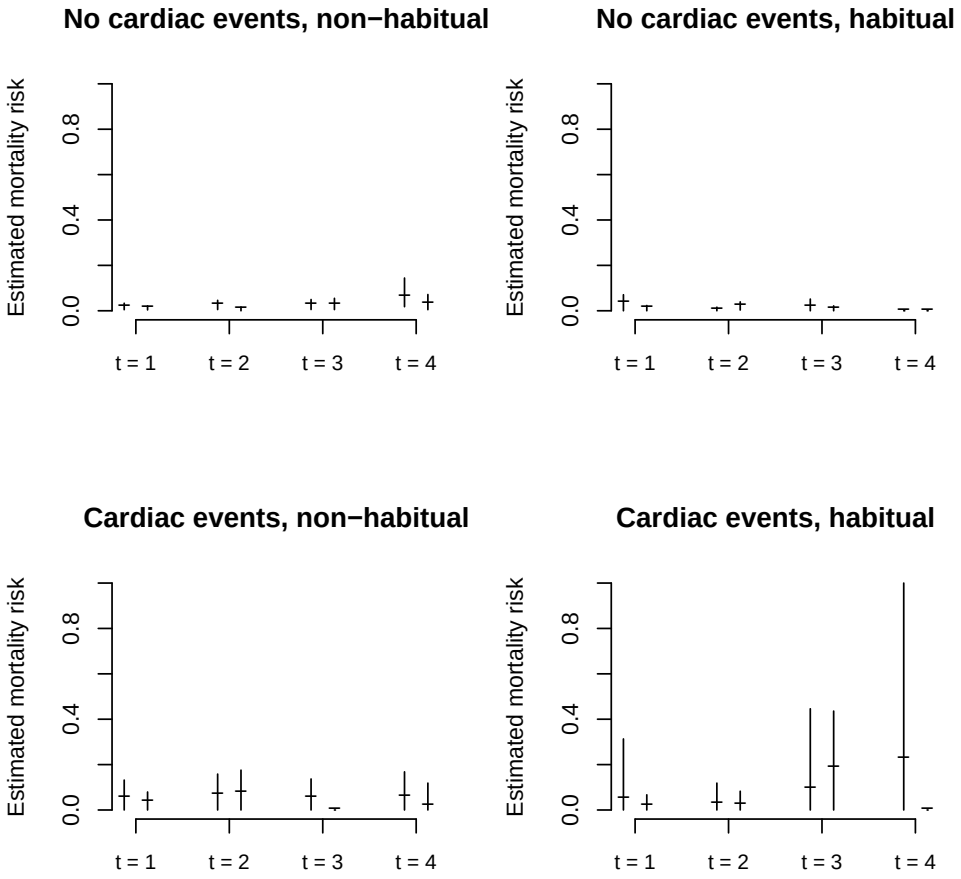

Figure 6: DR-IPCW mortaliy risk estimates with 95% bootstrap confidence intervals for age < 75. For each subpopulation and time point, the plot shows estimated counterfactual mortality risks under low LTPA on the left and those under high LTPA on the right.

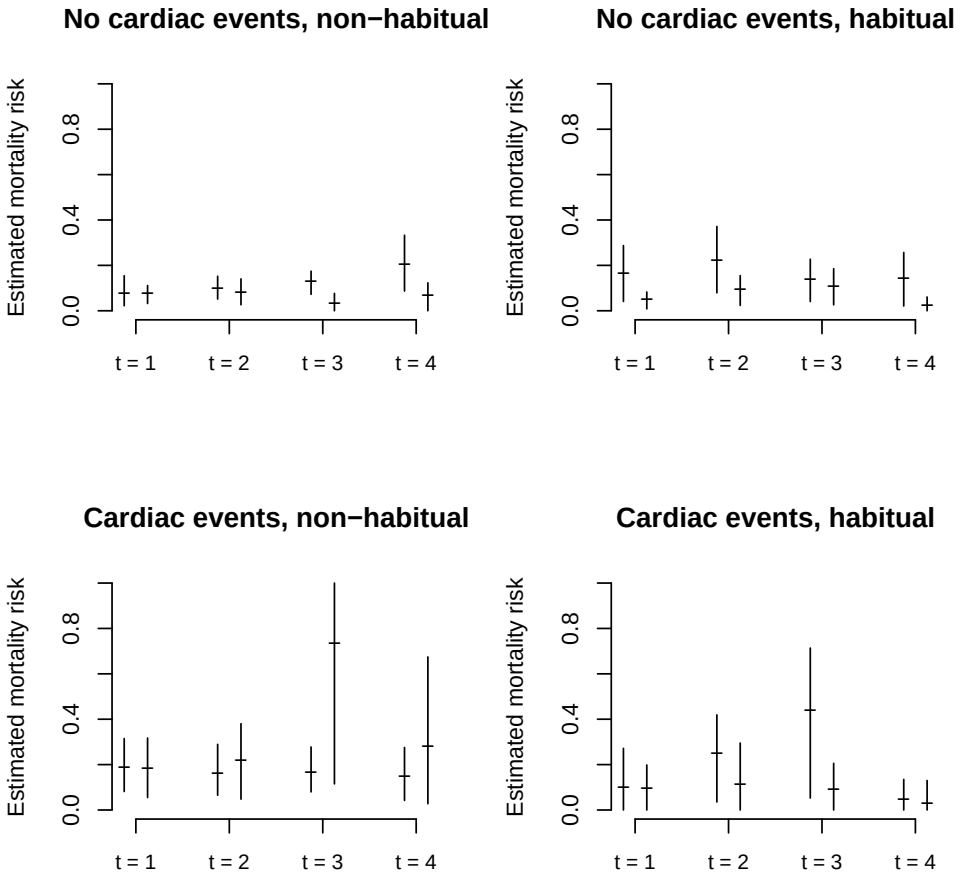

Figure 7: DR-IPCW mortality risk estimates with 95% bootstrap confidence intervals for age ≥ 75. For each subpopulation and time point, the plot shows estimated counterfactual mortality risks under low LTPA on the left and those under high LTPA on the right.

Table 5: Summary statistics for cohort of subjects of age $\geq$ 75 years with previous cardiac events and no history of habitual exercise. For each interview time point, the table summarizes the group of subjects in this cohort following the low-activity regimen and the group of subjects following the high-activity regimen.

	t=1		**t=2**		**t=3**		**t=4**	
	Low	High	Low	High	Low	High	Low	High
Mortality								
Deaths	9	6	9	5	10	8	10	4
Population	41	44	45	36	52	37	53	36
Crude mortality	22%	13.6%	20%	13.9%	19.2%	21.6%	18.9%	11.1%
Sex								
Female	58.5%	38.6%	64.4%	30.6%	65.4%	54.1%	71.7%	41.7%
Male	41.5%	61.4%	35.6%	69.4%	34.6%	45.9%	28.3%	58.3%
Age								
0th %ile	75	75	75	75	76	75	75	75
25th %ile	77	76	78	76	78	78	78	79
50th %ile	81	80	81	79	81	81	81	81
75th %ile	84	82	84	83	85	86	86	86
100th %ile	91	89	93	91	96	91	98	94
NRB								
0th %ile	–	–	0	0.19	0	0.23	0	0.17
25th %ile	–	–	0.31	0.72	0.48	0.71	0.46	0.72
50th %ile	–	–	0.65	0.92	0.74	0.84	0.67	0.9
75th %ile	–	–	0.9	1	0.88	1	0.84	1
100th %ile	–	–	1	1	1	1	1	1
Health								
Excellent health	4.9%	6.8%	4.4%	22.2%	5.8%	16.2%	11.3%	16.7%
Good health	43.9%	54.5%	37.8%	50%	57.7%	54.1%	58.5%	63.9%
Fair health	39%	27.3%	44.4%	27.8%	23.1%	21.6%	26.4%	16.7%
Poor health	12.2%	11.4%	13.3%	0%	13.5%	8.1%	3.8%	2.8%

6.6 Inference based on the influence curve

Figures 8 and 9 compare confidence intervals for the DR-IPCW estimates of the relative risk of mortality based on the bootstrap to those based on influence curve arguments. The latter confidence intervals were obtained through an application of the δ-method to the theoretical limiting distribution of $(\beta_n(1,v), \beta_n(0,v))$, finding confidence intervals first for $\log(\beta(1,v)/\beta(0,v))$ and then transforming them to the original scale since the distribution of $\log(\beta(1,v)/\beta(0,v))$ approaches the limiting normal distribution faster than does $\beta(1,v)/\beta(0,v)$.

The largest discrepancies between the two different types of confidence intervals are seen in the population of young subjects with a history of habitual exercise, where the influence-curve confidence intervals appear to greatly underestimate the variability of the point estimates. We note that the estimates in this population are among the most variable ones, suggesting that the amount of data they are based on may be too small for their distribution to be well approximated by an asymptotic limiting distribution. Hence it is not surprising that bootstrap confidence intervals perform better in these cases. For the remaining estimates, the influence-curve confidence intervals are generally not too much smaller than those based on the bootstrap. For estimates based on a sufficient amount of data, the potentially optimistic behavior of influence-curve confidence intervals resulting from data-adaptive estimation of the nuisance parameters thus appears to balance reasonably well with a slightly conservative behavior due to the use of an influence curve that has not yet been projected on the orthogonal complement of the nuisance tangent space.

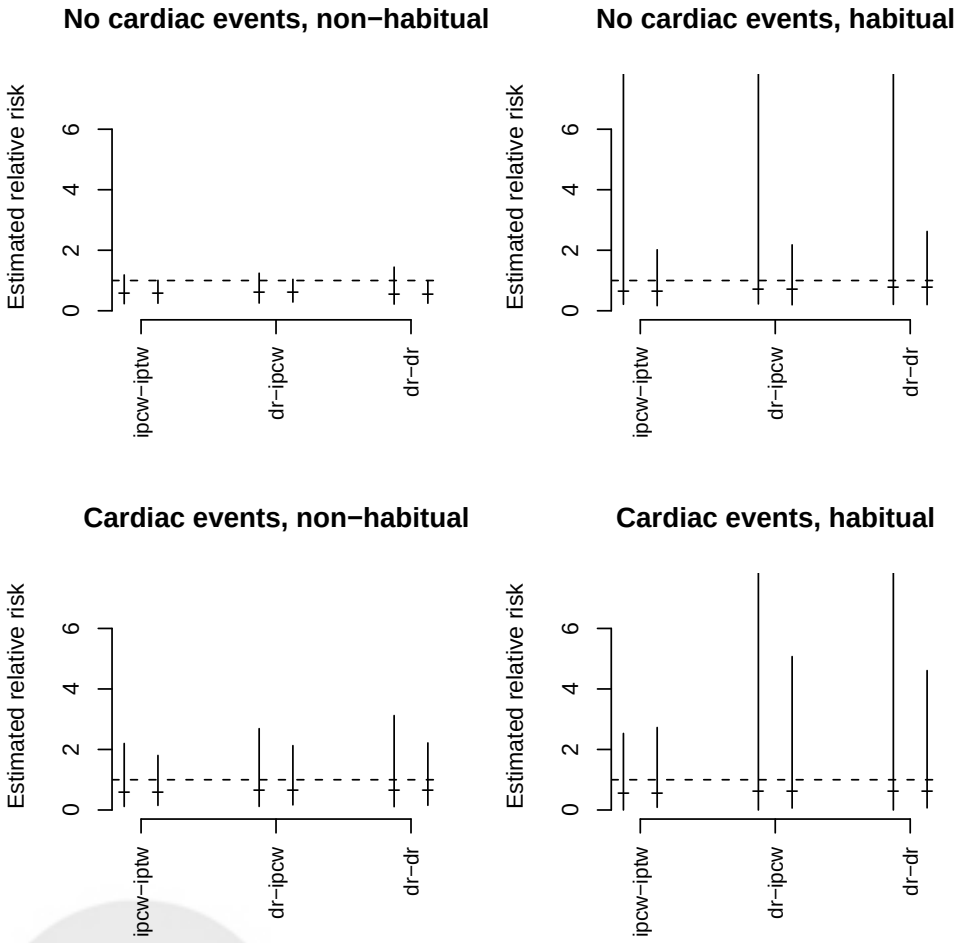

Figure 8: Comparison of inference for relative risk estimates for age < 75. For each subpopulation and estimator, the plot shows bootstrap confidence intervals on the left and those based on the influence curve on the right.

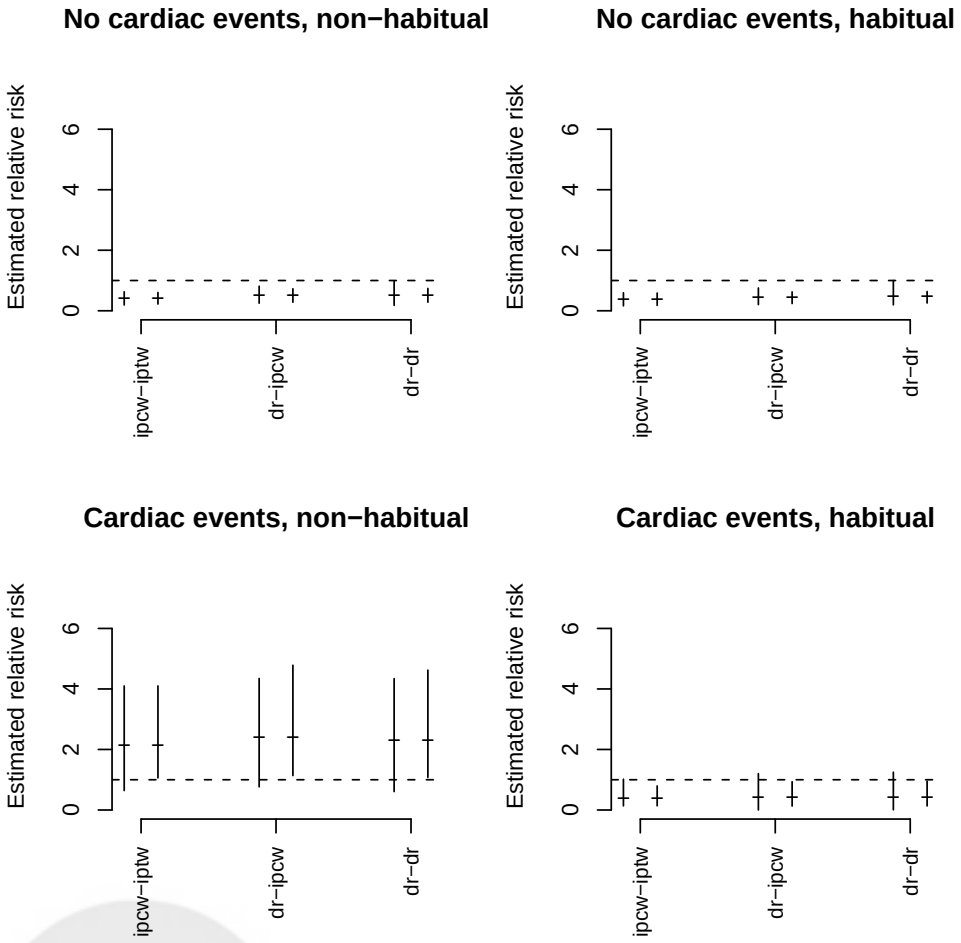

Figure 9: Comparison of inference for relative risk estimates for age ≥ 75. For each subpopulation and estimator, the plot shows bootstrap confidence intervals on the left and those based on the influence curve on the right.

7 Discussion

Our analyses suggest that high leisure-time physical activity confers a 40% reduction in two-year mortality risk among elderly people under the age of 75 years. Subjects above this age cut-off are estimated to experience a corresponding 60% reduction in risk unless they lack a history of habitual exercise and have previously had cardiac events. In that case, high leisure-time physical activity is in fact estimated to lead to a 2.33-fold increase in two-year mortality risk. While the precision of these causal effect estimates is limited among the younger subpopulation, the estimates approach statistical significance for older subjects who have previously experienced cardiac events and are in fact statistically significant for older subjects with previous cardiac events.

The DR-IPCW estimator that these results are based on relies on three major assumptions that need to be satisfied in order to obtain consistent estimates of causal effects from observational data. First, the temporal ordering assumption requires that covariates measured at time t are only affected by treatments at earlier time points. Since the treatment variable is measured over the course of one year, we had to define the covariates corresponding to time t as the covariates that were in fact measured at time $t-1$ in order to obtain a data structure that satisfies this assumption. We note that this time ordering problem is a wide-spread issue in longitudinal studies that is rarely addressed in analyzing the collected data.

Second, the sequential randomization assumption requires that our data set contain all relevant confounders of the relationship between leisure-time physical activity and mortality as well as all relevant confounders of the relationship between leisure-time physical activity and drop-out. Given the sizeable collection of variables available for capturing a subject's health status and past activity patterns, we are comfortable that this assumption is well approximated in the present case. Since the baseline covariates do not contain all the potential confounders available at later time points, the approximation may be slightly worse for the first time point. This is of interest since the risk increase reported for elderly subjects above the age of 75 who lack a history of habitual exercise and have experienced previous cardiac events is seen more strongly at the later three time points than at the first time point.

The estimator furthermore relies on an appropriate model for capturing the influence of these confounders on the drop-out process as well as at least one appropriate model for their impact on either the treatment variable or the risk of mortality. We are comfortable that the nuisance models used here satisfy this requirement since they are selected in a data-adaptive manner instead of being specified *a priori*. At the same time, we verified that the selected models make sense from a subject-matter point of view. We also note that most commonly used estimators would rely on correct specification of all three of these models as well as on correct specification of models for the conditional distribution of the covariate process given the past.

Third, the experimental treatment assignment assumption requires that there are essentially no realizations of the covariate process for which treatment or drop-out are assigned in a deterministic manner. In the absence of this absence, causal effects are not non-parametrically identifiable from observational data. While this assumption is rarely

evaluated in practice, we have employed a simulation-based approach for assessing the impact of any violation, theoretical or practical, of this assumption on the behavior of the DR-IPCW estimator. The results indicate that any such bias is negligible for those parameters that can be estimated with reasonable precision, with the small bias that is observed in such cases generally moving estimates closer to the null value.

As shown by Gill and Robins (2001) and Yu and van der Laan (2002), these assumptions place no restrictions on the data-generating distribution. Our analyses don't rely on any additional assumptions, allowing us thus to work in a non-parametric model. In particular, we do not require that corresponding mortality risks are identical across all four time points for our parameter of interest to be well-defined. At the same time there does not appear sufficient evidence against this assumption, which simplifies the interpretation of this parameter considerably. In examining the behavior of mortality risk estimates across different time points, we found that the risk increase reported for elderly subjects above the age of 75 who lack a history of habitual exercise and have experienced previous cardiac events is driven primarily by the estimates for the third time point. We verified, however, that this is unlikely to be due to problems with our adjustment for confounding and more likely related to the instability of the crude mortality estimates.

While inference for the DR-ICPW estimator would commonly be based on a boot-strap that ignores the variability introduced by selecting nuisance parameter models data-adaptively, the boot-strap procedure employed here fully takes this aspect into account. The reported confidence intervals are hence more honest and cannot be regarded as overly optimistic.

If we believe that the required assumptions are sufficiently satisfied for the DR-IPCW estimator to be consistent, we may turn to subject-matter considerations in an attempt to explain the results we report here. The reductions in mortality risk that we estimate to be experienced by the majority of subgroups as a consequence of high leisure-time physical activity are consistent with a substantial body of epidemiological research pointing to an association between physical activity among elderly people and reductions in cardiovascular morbidity and mortality (CDC, 1989; van Dam et al., 2002; Lee et al., 2003; Esposito et al., 2003; Rosano et al., 2005). The estimated 2.33-fold risk increase in elderly subjects above the age of 75 who lack a history of habitual exercise and who have previously experienced a cardiac event is surprising. The result could be due purely to chance, given some imprecision around the point estimate. Inspection of table 3 does not provide any obvious explanation. For this analysis, we did not evaluate the dates of the most recent cardiac events; therefore, we might speculate that early exercise as part of rehabilitation actually precipitated fatal outcomes. Nonetheless, on the whole, our results suggest that leisure-time physical activity does benefit those over 75 years.

Our analyses suggest that the benefits of high-level LTPA may be more pronounced in the older group of subjects. While this may be due to chance or to a real causal effect of LTPA on mortality, this result could also be due to the fact that the older group represents a robust survivor population derived from age cohorts with very different lifetime experiences. Some support for this idea can be found in Table 5 which shows that the percent of high

exercisers who rated their health as fair or poor was relatively stable over time and was lowest overall at the last two time points.

We propose five different estimators of the parameter of interest that are based on different combinations of the usual G-computation, inverse-weighting, and double robust approaches for the two layers of missingness corresponding to the treatment mechanism and right-censoring by drop-out. In particular, we propose a novel estimator that combines IPCW weights for dealing with potentially informative drop-out with a G-computation approach for dealing with the missingness corresponding to the treatment mechanism. While the five estimators agree fairly well in most cases, there are a few cases in which they produce somewhat conflicting estimates. In this paper, we used robustness considerations to justify focusing our attention on the DR-IPCW estimator. Another approach to dealing with this situation would be to apply the methodology of targeted maximum-likelihood estimation, recently introduced by van der Laan and Rubin (2006). By unifying estimating function methodology with likelihood-based estimation, this methodology provides G-computation, inverse-weighting, and double robust estimates that are algebraically identical, thus removing the need to reconcile any conflicting results.

References

CDC. Surgeon general's workshop on health promotion and aging: summary recommendations of physical fitness and exercise working group. *Journal of the American Medical Association*, 262:2507–2510, 1989.

K. Esposito, A. Pontillo, C. Di Palo, G. Giugliano, M. Masella, R. Marfella, and D. Giugliano. Effect of weight loss and lifestyle changes on vascular inflammatory markers in obese women: a randomized trial. *Journal of the American Medical Association*, 289(14):1799–1804, 2003.

R.D. Gill and J.M. Robins. Causal inference in complex longitudinal studies: continuous case. *The Annals of Statistics*, 29(6):1785–1811, 2001.

F.B. Hu, J.E. Manson, M.J. Stampfer, G. Colditz, S. Liu, C.G. Solomon, and W.C. Willett. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine*, 345:790–797, 2001.

D.R Jacobs, Jr. and M.A. Pereira. Physical activity, relative body weight, and risk of death among women (editorial). *New England Journal of Medicine*, 351:2753–2755, 2004.

C. Kooperberg, S. Bose, and C.J. Stone. Polychotomous regression. *Journal of the American Statistical Association*, 92:117–127, 1997.

W.E. Kraus, J.A. Houmard, B.D. Duscha, K.J. Knetzger, M.B. Wharton, J.S. McCartney, C.W. Bales, S. Henes, G.P. Samsa, J.D. Otvos, K.R. Kulkarni, and C.A. Slentz. Effects of the amount and intensity of exercise on plasma lipids. *New England Journal of Medicine*, 347:1483–1492, 2002.

I.M. Lee, H.D. Sesso, Y. Oguma, and R.S. Paffenberger, Jr. Relative intensity of physical activity and risk of coronary heart disease. *Circulation*, 107(8):1110–116, 2003.

R. Neugebauer and M.J. van der Laan. Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference*, 129:405–426, 2005a.

R. Neugebauer and M.J. van der Laan. Locally efficient estimation of nonparametric causal effects on mean outcomes in longitudinal studies. Technical Report 134, UC Berkeley Division of Biostatistics Working Paper Series, 2005b. URL `http://www.bepress.com/ucbbiostat/paper134`.

C. Rosano, E.M. Simonsick, T.B. Harris, S.B. Kritchevsky, J. Brach, M. Visser, K. Yaffe, and A.B. Newman. Association between physical and cognitive function in healthy elderly: the health, aging, and body composition study. *Neuroepidemiology*, 24(1-2):8–14, 2005.

P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

I.B. Tager, M. Hollenberg, and W. Satariano. Self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population. *American Journal of Epidemiology*, 147:921–931, 1998.

S.G. Thomas, D.A. Cunningham, P.A. Rechnitzer, A.P. Donner, and J.H. Howard. Determinants of the training response in elderly men. *Medicine and Science in Sports and Exercise*, 17:667–672, 1985.

R.M. van Dam, A.J. Schuit, E.J. Feskens, J.C. Seidell, and D. Kromhout. Physical activity and glucose tolerance in elderly men: the zutpen elderly study. *Medicine and Science in Sports and Exercise*, 34:1132–1136, 2002.

M.J. van der Laan and J.M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer Verlag, 2003.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006. URL `http://www.bepress.com/ijb/vol2/iss1/11`.

M.J. van der Laan, M.L. Petersen, and M.M. Joffe. History-Adjusted Marginal Structural Models and statically optimal dynamic treatment regimens. *The International Journal of Biostatitics*, 1(1):Article 4, 2005. URL `http://www.bepress.com/ijb/vol1/iss1/4`.

Y. Wang, M.L. Petersen, D. Bangsberg, and M.J. van der Laan. Diagnosing Bias in the Inverse-Probability-of-Treatment-Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. Technical Report 211, UC Berkeley Division of Biostatistics Working Paper Series, 2006. URL `http://www.bepress.com/ucbbiostat/paper211`.

Z. Yu and M.J. van der Laan. Construction of counterfactuals and the g-computation formula. Technical Report 122, UC Berkeley Division of Biostatistics Working Paper Series, 2002. URL `http://www.bepress.com/ucbbiostat/paper122`.

A Estimates

Table 6: Counterfactual mortality risk estimates with 95% bootstrap confidence intervals.

old	card	hab	ltpa	gcomp	gcomp.ipcw	iptw.ipcw	dr.ipcw	dr.dr
				0.03 (0.02, 0.04)	0.03 (0.02, 0.04)	0.03 (0.02, 0.05)	0.03 (0.02, 0.05)	0.03 (0.01, 0.05)
			✓	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.01 (0.01, 0.03)	0.02 (0.01, 0.03)	0.01 (0.01, 0.03)
		✓		0.01 (0.00, 0.03)	0.01 (0.00, 0.02)	0.02 (0.00, 0.03)	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)
		✓	✓	0.01 (0.01, 0.02)	0.01 (0.01, 0.02)	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)
	✓			0.05 (0.02, 0.08)	0.05 (0.02, 0.09)	0.06 (0.02, 0.12)	0.06 (0.01, 0.10)	0.05 (0.01, 0.10)
	✓		✓	0.03 (0.01, 0.06)	0.03 (0.01, 0.07)	0.03 (0.01, 0.06)	0.03 (0.01, 0.07)	0.03 (0.01, 0.07)
	✓	✓		0.10 (0.03, 0.18)	0.11 (0.03, 0.20)	0.10 (0.02, 0.25)	0.08 (0.00, 0.28)	0.08 (0.00, 0.28)
	✓	✓	✓	0.03 (0.00, 0.07)	0.03 (0.00, 0.07)	0.05 (0.00, 0.12)	0.05 (0.00, 0.10)	0.04 (0.00, 0.10)
✓				0.10 (0.08, 0.12)	0.11 (0.09, 0.14)	0.12 (0.09, 0.18)	0.12 (0.09, 0.17)	0.12 (0.08, 0.19)
✓			✓	0.06 (0.04, 0.08)	0.07 (0.05, 0.10)	0.05 (0.03, 0.06)	0.06 (0.03, 0.08)	0.06 (0.02, 0.09)
✓		✓		0.12 (0.09, 0.16)	0.13 (0.10, 0.17)	0.15 (0.10, 0.23)	0.16 (0.09, 0.22)	0.16 (0.08, 0.23)
✓		✓	✓	0.05 (0.03, 0.08)	0.07 (0.04, 0.10)	0.05 (0.03, 0.08)	0.06 (0.03, 0.09)	0.06 (0.03, 0.10)
✓	✓			0.14 (0.10, 0.18)	0.17 (0.13, 0.23)	0.16 (0.11, 0.23)	0.16 (0.11, 0.23)	0.16 (0.10, 0.23)
✓	✓		✓	0.12 (0.08, 0.18)	0.19 (0.12, 0.25)	0.34 (0.12, 0.61)	0.36 (0.13, 0.67)	0.35 (0.10, 0.66)
✓	✓	✓		0.12 (0.06, 0.17)	0.14 (0.07, 0.21)	0.22 (0.09, 0.34)	0.20 (0.07, 0.28)	0.20 (0.06, 0.28)
✓	✓	✓	✓	0.07 (0.04, 0.13)	0.10 (0.05, 0.18)	0.07 (0.03, 0.12)	0.07 (0.00, 0.14)	0.07 (0.00, 0.13)

Table 7: Relative risk estimates with 95% bootstrap confidence intervals.

old	card	hab	gcomp	gcomp.ipcw	iptw.ipcw	dr.ipcw	dr.dr
			0.54 (0.30, 1.02)	0.59 (0.31, 1.10)	0.50 (0.23, 1.18)	0.55 (0.25, 1.24)	0.50 (0.22, 1.44)
		✓	0.89 (0.33, 6.31)	0.95 (0.33, 8.91)	0.58 (0.21, 11.78)	0.66 (0.23, 12.57)	0.73 (0.21, 149.01)
	✓		0.66 (0.16, 2.49)	0.73 (0.18, 2.30)	0.52 (0.12, 2.20)	0.60 (0.12, 2.69)	0.59 (0.11, 3.12)
	✓	✓	0.31 (0.00, 1.41)	0.27 (0.00, 1.14)	0.49 (0.00, 2.53)	0.57 (0.00, Inf)	0.57 (0.00, Inf)
✓			0.60 (0.38, 0.91)	0.64 (0.42, 0.93)	0.37 (0.19, 0.63)	0.45 (0.25, 0.81)	0.46 (0.18, 0.99)
✓		✓	0.43 (0.26, 0.75)	0.53 (0.31, 0.84)	0.32 (0.16, 0.59)	0.39 (0.20, 0.75)	0.40 (0.20, 1.00)
✓	✓		0.86 (0.53, 1.44)	1.14 (0.64, 1.62)	2.08 (0.64, 4.10)	2.33 (0.76, 4.35)	2.23 (0.61, 4.34)
✓	✓	✓	0.60 (0.29, 1.58)	0.74 (0.30, 1.72)	0.33 (0.14, 1.02)	0.35 (0.00, 1.20)	0.36 (0.01, 1.25)

B Measured confounders

Variable	Definition
$FEMALE$	Sex, coded as an indicator variable for female
$VIG1$	Number of vigorous leisure-time physical activities participated in between the ages of 15 and 20
$VIG2$	Number of vigorous leisure-time physical activities participated in between the ages of 20 and 39
$VIG3$	Number of vigorous leisure-time physical activities participated in from age 40 up to the baseline interview
$MOD1$	Number of moderate leisure-time physical activities participated in between the ages of 15 and 20
$MOD2$	Number of moderate leisure-time physical activities participated in between the ages of 20 and 39
$MOD3$	Number of moderate leisure-time physical activities participated in from age 40 up to the baseline interview
DEC	An indicator of activity decline compared to 5 or 10 years ago
$HIGH$	An indicator of participation in high school sports
$HABITUAL$	An indicator of past habitual participation in vigorous leisure time physical activities
$ETSHM$	Years of exposure to environmental tobacco smoke at home prior to baseline
$ETSWK$	Years of exposure to environmental tobacco smoke at work prior to baseline

Table 8: Definition of the measured time-independent covariates that are considered as potential confounders.

Variable	Definition
$CARD$	An indicator for the previous occurrence of any of the following cardiac events: Angina, myocardial infarction, congestive heart failure, coronary by-pass surgery, and coronary angioplasty
$CHRON$	An indicator for the presence of any of the following chronic health conditions: stroke, cancer, liver disease, kidney disease, Parkinson's disease, and diabetes mellitus
HLT	Current self-rated health (excellent, good, fair, poor)
$DEPSCORE$	A depression score based on a questionnaire about the emotional state during the past week
$ANTIDEP$	An indicator for current use of antidepressant medication
NRB	A summary measure of physical functioning over the past month
BMI	Current body mass index, summarized into an ordinal variable with three categories
SMK	Smoking status (current, ex, never)
$LIVAR$	Living arrangement (alone, with spouse, with non-spouse)
AGE	Age in years

Table 9: Definition of the measured time-dependent covariates that are considered as potential confounders.

C Covariate imputation

Variable	Imputation Method	Frequency	Percent
DEC	Not missing	2,084	99.62
	Typical value	8	0.38
$ETSHM$	Not missing	2,064	98.66
	Typical value	28	1.34
$ETSHM$	Not missing	2,056	98.28
	Typical value	36	1.72
MOD	Not missing	2,091	99.95
	Typical value	1	0.05
VIG	Not missing	2,091	99.95
	Typical value	1	0.05

Table 10: Among the time-independent covariates, $FEMALE$, $HIGH$ and $HABITUAL$ have no missing values. The table above summarizes how the remaining time-independent covariates were imputed.

Variable	Imputation Method	Frequency	Percent
HLT	Not missing	7,248	99.66
	Past measurement	7	0.10
	Future measurement	16	0.22
	Typical value	2	0.03
$CHRONIC$	Not missing	7,196	98.94
	Past measurement	45	0.62
	Future measurement	24	0.33
	Typical value	8	0.11
BMI	Not missing	5,021	96.91
	Past measurement	63	1.22
	Future measurement	17	0.33
	Typical value	80	1.54
NRB	Not missing	5,065	97.76
	Past measurement	106	2.05
	Future measurement	6	0.12
	Typical value	4	0.08
$DEPSCORE$	Not missing	4,852	93.65
	Past measurement	155	2.99
	Future measurement	61	1.18
	Typical value	113	2.18
$LIVAR$	Not missing	5,166	99.71
	Past measurement	15	0.29
	Future measurement	0	0.00
	Typical value	0	0.00

Table 11: Among the time-dependent covariates, $CARD$, AGE, SMK, and $ANTIDEP$ have no missing values. The table above summarizes how the remaining time-dependent covariates were imputed.

D Nuisance parameter models

The following models for the treatment mechanism were selected by `polyclass()` for the four different time points. The variable `ltpa.yr.t` refers to the original LTPA measurement in METs for time point t. Note that coefficient estimates give the log odds-ratio for low LTPA compared to high LTPA rather than the other way around.

Table 12: Treatment model for t = 1 based on 2074 observations and 24 variables.

var1	knot1	var2	knot2	beta1
NA		NA		-0.705
mod3		NA		-0.580
decline		NA		0.368
vig3		NA		-0.675
vig1		NA		-0.119
vig3		mod3		0.101
age		NA		0.021
female		NA		0.344
decline		mod3		0.204

Table 13: Treatment model for t = 2 based on 1748 observations and 44 variables.

var1	knot1	var2	knot2	beta1
NA				2.992
lpta.yr.1				-0.057
lpta.yr.1	57			0.045
nrb.1				-1.408

Table 14: Treatment model for t = 3 based on 1356 observations and 67 variables.

var1	knot1	var2	knot2	beta1
NA				2.916
lpta.yr.2				-0.074
lpta.yr.1				-0.019
nrb.1				-1.866
lpta.yr.2	19.5			0.044
bmi.1				0.041

Table 15: Treatment model for t = 4 based on 1098 observations and 90 variables.

var1	knot1	var2	knot2	beta1
NA				-0.719
lpta.yr.3				-0.081
lpta.yr.3	33			0.071
lpta.yr.2				-0.013
age				0.049
lpta.yr.1				-0.010

The following models for the drop-out mechanism were selected by `polyclass()` for the four different time points. Note that coefficient estimates give the log odds-ratio for remaining in the study compared to dropping out of the study rather than the other way around.

Table 16: Drop-out model for t = 2 based on 1985 observations and 43 variables.

var1	knot1	var2	knot2	beta1
NA				3.961
age				-0.028

Table 17: Drop-out model for t = 3 based on 1609 observations and 66 variables.

var1	knot1	var2	knot2	beta1
				1.679

Table 18: Drop-out model for t = 4 based on 1280 observations and 89 variables.

var1	knot1	var2	knot2	beta1
NA				5.309
age				-0.050
smoke2.2				-0.544
chronic.3				0.399
antidep.3				2.050

The following models for the the two-year risk of mortality were obtained for the four different time points by the two-step procedure described in section 5.6. Note that coefficient estimates

give the log odds-ratio for dying within the following two years compared to surviving the following two years rather than the other way around.

Table 19: Failure model for t = 1 based on 2074 observations and 21 variables.

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.436	0.935	-3.675	0.000
imp.etshome	1.733	0.485	3.572	0.000
decline	0.612	0.261	2.346	0.019
A	-0.546	0.566	-0.964	0.335
binage	0.239	1.165	0.205	0.838
card	0.174	1.858	0.094	0.925
habitual	0.271	1.775	0.152	0.879
A:binage	0.625	0.722	0.866	0.387
A:card	0.394	1.110	0.355	0.723
binage:card	1.431	2.187	0.654	0.513
A:habitual	-0.282	1.042	-0.271	0.787
binage:habitual	1.899	2.107	0.901	0.368
card:habitual	4.200	2.927	1.435	0.151
A:binage:card	-0.808	1.336	-0.605	0.545
A:binage:habitual	-1.089	1.281	-0.850	0.395
A:card:habitual	-2.379	1.864	-1.276	0.202
binage:card:habitual	-6.437	3.579	-1.799	0.072
A:binage:card:habitual	3.602	2.281	1.579	0.114

Table 20: Failure model for t = 2 based on 1748 observations and 41 variables.

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.474	1.120	-1.315	0.188
hlt4.1	1.253	0.409	3.063	0.002
nrb.1	-0.846	0.422	-2.007	0.045
etswork	-0.484	0.189	-2.559	0.010
etswork 2	0.519	0.196	2.653	0.008
imp.etswork	1.564	0.581	2.691	0.007
A	-1.396	0.808	-1.728	0.084
binage	0.475	1.204	0.395	0.693
card	-1.032	1.945	-0.531	0.595
habitual	-3.403	2.315	-1.470	0.142
A:binage	0.905	0.911	0.993	0.321
A:card	1.438	1.311	1.097	0.273
binage:card	0.350	2.224	0.157	0.875
A:habitual	2.292	1.368	1.675	0.094
binage:habitual	3.583	2.507	1.429	0.153
card:habitual	4.227	3.655	1.157	0.247
A:binage:card	-0.682	1.512	-0.451	0.652
A:binage:habitual	-2.258	1.518	-1.487	0.137
A:card:habitual	-2.892	2.245	-1.288	0.198
binage:card:habitual	-3.240	4.088	-0.793	0.428
A:binage:card:habitual	2.328	2.539	0.917	0.359

Table 21: Failure model for t = 3 based on 1356 observations and 64 variables.

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.547	1.028	-1.504	0.133
nrb.1	-1.235	0.609	-2.029	0.042
imp.bmi.2	1.705	0.478	3.569	0.000
chronic.1	0.634	0.245	2.584	0.010
nrb.2	0.100	0.766	0.131	0.896
nrb.20.77	-5.108	1.992	-2.564	0.010
female	-0.773	0.259	-2.988	0.003
A	-0.044	0.609	-0.073	0.942
binage	2.424	1.187	2.042	0.041
card	14.655	711.080	0.021	0.984
habitual	-1.092	2.329	-0.469	0.639
A:binage	-1.462	0.882	-1.658	0.097
A:card	-14.443	711.078	-0.020	0.984
binage:card	-16.752	711.081	-0.024	0.981
A:habitual	0.169	1.376	0.123	0.902
binage:habitual	-0.278	2.580	-0.108	0.914
card:habitual	-13.565	711.087	-0.019	0.985
A:binage:card	16.607	711.079	0.023	0.981
A:binage:habitual	1.158	1.619	0.715	0.475
A:card:habitual	14.690	711.081	0.021	0.984
binage:card:habitual	16.494	711.090	0.023	0.981
A:binage:card:habitual	-17.250	711.082	-0.024	0.981

Table 22: Failure model for t = 4 based on 1098 observations and 87 variables.

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.950	0.978	-0.971	0.331
imp.nrb.3	1.370	0.420	3.258	0.001
nrb.3	-1.896	0.482	-3.936	0.000
A	-0.553	0.642	-0.862	0.389
binage	0.413	1.139	0.363	0.717
card	-1.000	2.478	-0.404	0.686
habitual	-17.313	3093.055	-0.006	0.996
A:binage	0.283	0.798	0.354	0.723
A:card	0.685	1.586	0.432	0.666
binage:card	0.994	2.718	0.366	0.715
A:habitual	0.757	1738.354	0.000	1.000
binage:habitual	18.731	3093.056	0.006	0.995
card:habitual	34.559	4433.056	0.008	0.994
A:binage:card	-0.420	1.779	-0.236	0.813
A:binage:habitual	-2.272	1738.354	-0.001	0.999
A:card:habitual	-17.590	3620.340	-0.005	0.996
binage:card:habitual	-37.598	4433.057	-0.008	0.993
A:binage:card:habitual	19.196	3620.341	0.005	0.996