BIOMETRICS 000, 1–26 DOI: 000

April 2020

Weighted Regression Analysis to Correct for Informative Monitoring Times and Confounders in Longitudinal Studies

Janie Coulombe*, Erica E. M. Moodie, and Robert W. Platt*****

Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal H3A 1A2, Canada.

**email:* janie.coulombe@mail.mcgill.ca

***email:* erica.moodie@mcgill.ca

****email:* robert.platt@mcgill.ca

SUMMARY: We address estimation of the marginal effect of a time-varying binary treatment on a continuous longitudinal outcome in the context of observational studies using electronic health records, when the relationship of interest is confounded, mediated and further distorted by an informative visit process. We allow the longitudinal outcome to be recorded only sporadically and assume that its monitoring timing is informed by patients' characteristics. We propose two novel estimators based on linear models for the mean outcome that incorporate an adjustment for confounding and informative monitoring process through generalized inverse probability of treatment weights and a proportional intensity model respectively. We allow for a flexible modelling of the intercept function as a function of time. Our estimators have closed-form solutions, and their asymptotic distributions can be derived. Extensive simulation studies show that both estimators outperform standard methods such as the ordinary least squares estimator or estimators that only account for informative monitoring or confounders. We illustrate our methods using data from the *Add Health* study, assessing the effect of depressive mood on weight in adolescents.

KEY WORDS: Confounding bias; Informative monitoring times; Longitudinal data; Marginal effect of treatment.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Consider a setting where we are interested in understanding the cross-sectional impact of an exposure on an outcome, for example a physician is interested in the impact of a current clinical measurement on the current presence of an illness, and their patients are repeatedly assessed over time. To learn about such an association, we may turn to electronic health records (EHRs). Longitudinal outcomes, particularly those drawn from EHR data, may be measured irregularly across patients. The time points at which they are recorded may depend on patients' health condition, which may on its turn be linked with the value of the outcome measured at those visit times, leading to imbalances in the data similar to those observed in selection bias. Moreover, confounders and mediating variables (Greenland and Robins, 1986) occur simultaneously with informative monitoring times in most observational studies, and thus must also be accounted for.

In this work, we focus on making inference on the marginal effect of a binary, time-varying treatment on a continuous outcome, measured repeatedly over time. To model longitudinal outcomes in contexts where monitoring times are irregular or informative, several methods have been proposed, but none of them simultaneously considered confounding in a setting with continuous outcomes. When monitoring times are informative, Robins et al. (1995) proposed a weighted extension of the generalized estimating equations of Liang and Zeger (1986) to estimate the marginal effect of intervention on a longitudinal outcome. In their method, inverse probability of response weights were incorporated into estimating equations to adjust for nonrandom missingness, which addressed the problem of informative monitoring times but was restricted to the case where there is a common set of monitoring times for all individuals, which is often not the case in observational studies. In 2001, Lin and Ying (2001) developed a class of closed-form estimators for the marginal effect of variables on the mean outcome that accounted for informative monitoring times and allowed for those times to vary across individuals. Several innovations followed, which we detail further in the following section, many of which are covered in the review of Pullenayegum and Lim (2016).

We extend the existing literature further and propose two new and flexible estimators for the marginal effect of a (potentially time-varying) binary treatment on a longitudinal continuous outcome for settings in which the exposure is not randomized. In our methods, we allow the mediators, the exposure and other covariates to affect the timing of the outcome monitoring, and both the confounders and the mediators to vary in time. The first estimator is a semiparametric extension that builds on the work of Bůžková and Lumley (2009). The second estimator is a weighted least squares type estimator that incorporates two time-varying weights. This latter flexible estimator provides a simpler and more intuitive alternative to the first, with comparable performance. Its asymptotic variance is derived. In simulation studies, we compare both estimators and more standard ones in different contexts of dependency between covariates and monitoring times.

The remainder of this article is organized as follows: Section 2 introduces the notation, assumptions and inference procedure. Section 3 presents the details of the simulation studies and the results. Section 4 applies the methodology to the analysis of the data from the *Add Health* study (Harris et al., 2009a). Finally, we provide some concluding remarks in Section 5.

2. Methods

2.1 *Background*

Lin and Ying (2001) considered the following marginal model:

$$E\left[Y_i(t)|\mathbf{X_i}(\mathbf{t})\right] = \alpha(t) + \boldsymbol{\beta}'\mathbf{X_i}(\mathbf{t}), \tag{1}$$

with $\alpha(t)$ an arbitrary function of time t, $\mathbf{X}(\mathbf{t})$ a design matrix and $Y_i(t)$ a continuous longitudinal outcome. They assumed a proportional intensity model for the monitoring times of the outcome, which monitoring times were only allowed to depend on covariates in the outcome model, $\mathbf{X}(\mathbf{t})$. They proposed a semiparametric estimator for $\boldsymbol{\beta}$ in (1), which does not require estimation of the intercept $\alpha(t)$. In 2009, Bůžková and Lumley (2009) proposed to incorporate a weight in Lin and Ying's estimator that accounts for the dependency between monitoring times and any covariates that are *not* in the design matrix $\mathbf{X}(\mathbf{t})$. In particular, their approach allows for any mediators of the treatment-outcome relationship to affect monitoring times.

Tan et al. (2014) presented a summary of some of the extensions of Lin and Ying's estimator and proposed a few developments of existing methods. Other authors have proposed fully parametric methods to jointly model the visit and outcome processes (Lipsitz et al., 2002; Ryu et al., 2007), or introduced shared latent effects to link the outcome and the visit processes (e.g., Sun et al., 2012; Cai et al., 2012). Most recently, Zhu et al. (2017) proposed an estimator for interval-censored outcomes when confounding and irregular visit times may be present. They were among the first to consider these two features, however the method is focused on a very particular outcome type.

The problem of accounting for mediators and confounding variables in observational studies has been addressed via several methods. It is now well-known that mediating variables should not be included in the design matrix of the outcome model if the estimand is the total effect of exposure on the outcome (Rosenbaum, 1985). Propensity score methods such as inverse probability of treatment (IPT) weights are commonly used to adjust for imbalances across treatment groups due to confounders (Rosenbaum and Rubin, 1983; Robins et al., 2000a). The standard IPT weight for a binary and time-fixed treatment I_i, baseline confounders $\mathbf{K_i}$ and parameters $\boldsymbol{\omega}$ is given by

$$\mathrm{e}_i(\boldsymbol{\omega}) = \frac{1}{\mathbb{I}_{(I_i=1)} P(I_i = 1|\mathbf{K_i};\boldsymbol{\omega}) + \mathbb{I}_{(I_i=0)}(1 - P(I_i = 1|\mathbf{K_i};\boldsymbol{\omega}))}, \tag{2}$$

where $\mathbb{I}_{(I_i=1)}$ is an indicator function for treatment $I_i = 1$. We typically estimate the parameters $\boldsymbol{\omega}$ by fitting a logistic regression model with dependent variable I_i and predictors $\mathbf{K_i}$.

2.2 *Assumptions*

Suppose that we have a random sample of n individuals, indexed by $i = 1, ..., n$. We are interested in the marginal effect of a binary intervention $I_i(t)$ on the longitudinal, continuous outcome $Y_i(t)$, where t represents the time. We use the Neyman-Rubin potential outcome framework (Neyman, 1923; Rubin, 1974) to express the estimand of interest, which is the causal contrast $E\left[Y_{i1}(t) - Y_{i0}(t)\right]$ where $Y_{i1}(t)$ corresponds to the outcome that would have been observed at time t, had individual i been treated by $I_i(t) = 1$, and $Y_{i0}(t)$, had the individual been treated by $I_i(t) = 0$. More specifically, we focus on a time-invariant marginal effect of the exposure on the outcome recorded at the same monitoring time.

The terms *treatment*, *intervention* and *exposure* are used interchangeably to refer to $I_i(t)$, and *monitoring times* and *visit times* refer to the times when the outcome $Y_i(t)$ is observed. We use bold notation to refer to vectors and matrices. We now detail the assumptions required about the outcome model (O1-O2), the visit process (V1-V2), the treatment model (P1-P3), and the total follow-up time (C1); these assumptions are required for consistency of our proposed estimators.

To estimate $E\left[Y_{i1}(t) - Y_{i0}(t)\right]$, one can also estimate the contrast $E\left[Y_i(t)|I_i(t) = 1\right] - E\left[Y_i(t)|I_i(t) = 0\right]$ in a pseudo-population where there is no imbalance in covariates between treatment groups due to confounding and the monitoring process. In a setting with no such imbalance, we assume that treatment groups are similar in their characteristics and that patient groups are interchangeable prior to exposure. The following marginal linear model for the mean can be used for estimation:

$$E[Y_i(t)|I_i(t)] = \alpha(t) + \beta I_i(t). \tag{O1}$$

The parameter β in (O1) is exactly equal to $E\left[Y_i(t)|I_i(t) = 1\right] - E\left[Y_i(t)|I_i(t) = 0\right]$ so it might represent a valid estimate for the causal contrast of interest. However, we are aware in our setting of an underlying confounding process, and the following conditional model for the mean is a sensible model to use to estimate the conditional effect of treatment β_I:

$$E[Y_i(t)|I_i(t), \mathbf{K_i(t)}] = \alpha(t) + \beta_I I_i(t) + \boldsymbol{\beta_K}'\mathbf{K_i(t)} \tag{O2}$$

for $\mathbf{K_i(t)}$ the confounders of the relationship $(I_i(t), Y_i(t))$. Depending on the distribution of confounders $\mathbf{K_i(t)}$ in the sample under study, an estimator of β in the marginal model in (O1) might be biased for $E\left[Y_{i1}(t) - Y_{i0}(t)\right]$, due to imbalances in the confounders between treatment groups. Moreover, the model in (O2) is marginalized over other covariates not included in $I_i(t)$ or $\mathbf{K_i(t)}$ that may affect both the outcome and the monitoring times. For now, we do not consider explicit modeling of the covariates affecting monitoring times which could bias an estimate for β_I in (O2), and focus on the conditional model (O2). We see later how we can obtain an estimate for the average treatment effect in a pseudo-population where there are no imbalances between treatment groups with respect to $\mathbf{K_i(t)}$, and no imbalances in observed/unobserved outcomes due to an informative monitoring process.

Let the intercept $\alpha(t)$ remain unspecified in (O2). In addition to confounding, assume that the relationship between $I_i(t)$ and $Y_i(t)$ may be mediated by a vector of (potentially time-varying) covariates $\mathbf{Z_i(t)}$ which are in the causal path from the exposure $I_i(t)$ to the outcome $Y_i(t)$; see Bůžková and Lumley (2005) for an asthma-related example.

Assume that the longitudinal outcome $Y_i(t)$ is only observed at times $T_{i1}, ..., T_{iK_i}$, with $N_i(t) = \sum_{k=1}^{K_i} I(T_{ik} \leq t)$. Note that other patient features might be recorded and available in between times when the outcome is recorded. $N_i(t)$ is used to denote the number of monitoring times by time t, for individual i. The quantity $dN_i(t)$ is equal to 1 if $Y_i(t)$ is measured at time t and 0 otherwise, and τ represents the maximum follow-up time in the cohort under study.

We suppose that the relationship between $I_i(t)$ and $Y_i(t)$ may be distorted by an informative monitoring process, and that monitoring at time t depends on the set of covariates $\mathbf{V_i(t)} = \{\mathbf{Z_i(t)}, I_i(t)\}$ through a proportional intensity model for the monitoring times:

$$E[dN_i(t)|\mathbf{V_i(t)}] = \xi_i(t) \exp\left(\gamma_V' \mathbf{V_i(t)}\right) d\Lambda_0(t), \tag{V1}$$

where the function $\Lambda_0(\cdot)$ is arbitrary and nondecreasing, and $\xi_i(t)$ is the indicator of individual i still being in the study at time t. We assume that for each time $0 < t < C_i$, for a certain time granularity (e.g. daily), and for each individual i, we have $0 < P[dN_i(t)|\mathbf{V_i(t)}] < 1$. We restrict the assumption to a particular granularity, as positivity is unlikely to hold when time is continuous.

Suppose that $\mathbf{V_i(t)}$ contains all common predictors of the monitoring times and the outcome,

$$N_i(t) \perp Y_i(t)|\mathbf{V_i(t)}. \tag{V2}$$

In fact, note that $\mathbf{Z(t)} \subset \mathbf{V(t)}$ may contain any mediator of the relationship between $I_i(t)$ and $Y_i(t)$, but also any other variable that is not the intervention but that affects monitoring times.

Note that the modelling of monitoring times through equation (V1) requires all covariates affecting monitoring times to be available at any time t, $0 \leq t < C_i$, $\forall i$ during follow-up (again, under a particular time granularity, e.g. daily). We note that administrative databases or EHRs often contain the information on drugs prescribed or previous diagnostics at any time (even in between times when the outcome is recorded) and these values can be carried forward in between monitoring times so as to use as much information as possible. In clinical practice, in the absence of new measurements, this information may be relied on to make decisions (Cao et al., 2016).

For the exposure, we assume conditional exchangeability, stable-unit treatment value and positivity of treatment, which respectively correspond to:

$$I_i(t) \perp \{Y_{i0}(t), Y_{i1}(t)\} \,|\mathbf{K_i}(\mathbf{t}) \tag{P1}$$

$$\{Y_{i0}(t), Y_{i1}(t)\} \,|I_i(t) = \{Y_{i0}(t), Y_{i1}(t)\} \,|I'_i(t) \text{ if } I_i(t) = I'_i(t) \tag{P2}$$

$$0 < P(I_i(t) = 1|\mathbf{K_i}(\mathbf{t})), P(I_i(t) = 0|\mathbf{K_i}(\mathbf{t})) < 1. \tag{P3}$$

These conditions are necessary to use propensity score methods to adjust for confounding, along with correct model specifications.

While the maximum follow-up time is τ, it may be that some individuals are not followed after a certain point. Let C_i denote the end of follow-up ("censoring" time, though we are not working in a time-to-event context) for individual *i*; we consider that the end of follow-up is administrative and non-informative, that is

$$E\left[Y_i(t)|I_i(t), \mathbf{K_i}(\mathbf{t}), C_i \geq t\right] = E\left[Y_i(t)|I_i(t), \mathbf{K_i}(\mathbf{t})\right]. \tag{C1}$$

We note that this assumption could be circumvented by using inverse probability of censoring weights to adjust for informative dropout. See, for instance, Robins et al. (2000a).

The causal diagram in Figure 1, panel A depicts the structure of the data generating mechanism at time t. Panel B shows the presumed underlying data mechanism for our analysis of the *Add Health* study, presented in Section 4. Note in Figure 1 that we assume that the confounders and the mediators vary in time, which is allowed but is not necessary. Even when these variables vary in time, their effects on the monitoring times are assumed constant over time (i.e. we estimate $\boldsymbol{\gamma}_V$ rather than $\boldsymbol{\gamma}_V(t)$). Finally, we note that knowledge about the problem under study should inform the best choice for the set $\mathbf{K_i(t)}$, which may incorporate covariates measured at time t, as well as at previous time s, for $s < t$. Settings with time-dependent confounding are allowed, as long as the set of confounders include all potential confounders of the marginal relationship under study at a given time and that mediators are not conditioned upon in the outcome model.

[Figure 1 about here.]

Dotted arrows in Figure 1 refer to potential relationships we may want to consider.

2.3 *Existing methods*

Lin and Ying (2001) proposed a semiparametric estimator for β in the marginal model (O1) without reference to a particular covariate or intervention of interest. Their method did not account for the variables that affected the monitoring times whenever those variables were not contained in the design matrix for the outcome model. Bůžková and Lumley (2009) extended their work to account for those other variables. They built an estimator for the marginal effect of treatment based on the stochastic process $M_i(t;\beta,\boldsymbol{\gamma_V},\mathscr{A})$ which, in our case, is defined by

$$M_i(t;\beta,\boldsymbol{\gamma_V},\mathscr{A}) = \int_0^t \left(Y_i(s) - \beta I_i(s)\right) dN_i(s) - \xi_i(s)\exp\left(\boldsymbol{\gamma_V}'\mathbf{V_i}(\mathbf{s})\right) d\mathscr{A}(s), \tag{3}$$

where $\mathscr{A}(t) = \int_0^t \alpha(s)d\Lambda(s)$. They defined a weighted version $G_i(t;\beta,\boldsymbol{\gamma},\mathscr{A})$ of that process, with

$$G_i(t;\beta,\boldsymbol{\gamma},\mathscr{A}) = \int_0^t \frac{1}{\rho_i(s;\boldsymbol{\gamma})} dM_i(s;\beta,\boldsymbol{\gamma_V},\mathscr{A}) \tag{4}$$

with the stabilized rate ratio weight $\rho_i(s;\boldsymbol{\gamma})$, given in our setting by

$$\rho_i(s;\boldsymbol{\gamma}) = \frac{\exp\left(\boldsymbol{\gamma_1}'\mathbf{Z_i}(\mathbf{s}) + \gamma_2 I_i(s)\right)}{\exp\left(\gamma_I I_i(s)\right)}. \tag{5}$$

Note that $\gamma_{\mathbf{1}}'\mathbf{Z_i}(\mathbf{s}) + \gamma_2 I_i(s) = \boldsymbol{\gamma_V}'\mathbf{V_i}(\mathbf{s})$. The weight (5) allows their estimator to consider the dependency between $\mathbf{Z(t)}$ and the monitoring times while not adding $\mathbf{Z(t)}$ directly into the design matrix. It also accounts for the dependency between the covariates in the design matrix of the outcome model (here, $I_i(t)$) and the monitoring times. The parameters $\boldsymbol{\gamma_1}$ and γ_2 in (5) can be estimated by fitting a proportional intensity model for monitoring times with $\mathbf{Z}(\mathbf{t})$ and $\mathbf{I}(\mathbf{t})$ as covariates, while γ_I is estimated using the same type of model with only $\mathbf{I}(\mathbf{t})$ as a covariate.

Bůžková and Lumley (2009) show that $E\left[dG_i(t;\beta,\boldsymbol{\gamma},\mathscr{A})|I_i(t)\right]=0$ under assumptions (O1), (C1), (V1) and (V2). They further build estimating equations for β in (O1). In our setting where the design matrix is $\mathbf{I}(\mathbf{t})$, their procedure yields the following estimator:

$$\begin{aligned}\widehat{\beta}_{BL}=&\left[\sum_{i=1}^{n}\int_0^{\tau}\frac{W(t)}{\rho_i(t;\boldsymbol{\gamma})}\left(I_i(t)-\overline{I}(t;\gamma_I)\right)^2 dN_i(t)\right]^{-1}\\ &\times\sum_{i=1}^{n}\int_0^{\tau}\frac{W(t)}{\rho_i(t;\boldsymbol{\gamma})}\left(I_i(t)-\overline{I}(t;\gamma_I)\right)\left(Y_i(t)-\overline{Y}^{*}(t;\gamma_I)\right)dN_i(t),\end{aligned}\qquad(6)$$

which is a least squares type estimator where the design matrix is the vector $\left(\mathbf{I}(\mathbf{t})-\overline{\mathbf{I}}(\mathbf{t};\gamma_{\mathbf{I}})\right)$, the outcome vector is given by $\left(\mathbf{Y}(\mathbf{t})-\overline{\mathbf{Y}}^{*}(\mathbf{t};\gamma_{\mathbf{I}})\right)$, $W(t)$ is an arbitrary time-dependent weight that may be used to reduce the variance, and $\overline{Y}^{*}(t;\gamma_I)$ a weighted average of the nearest-neighbor approximation to Y at time t (which is also used to reduce the variance of the estimator). The re-centering of $I_i(t)$ by its adjusted mean in (6) eliminates from the estimation the intercept $\alpha(t)$ in (O1) and avoids having to model the relationship between the mean outcome and time t, hence the semiparametric and more flexible nature of the estimator. The estimating equations that Bůžková and Lumley (2009) used are sums of independent zero-mean random vectors, and the variance of their estimator can be derived using standard asymptotic theory along with Taylor expansions. In what follows, we use $W(t)=1\ \forall t$.

In order to estimate the adjusted means, the proportional intensity model (V1) is fitted with only the predictor $I_i(t)$. The coefficient $\widehat{\gamma}_I$ for $I_i(t)$ is used to compute the weighted means. For any vector $\mathbf{R}(\mathbf{t})$ in general, we have:

$$\overline{R}(t;\widehat{\gamma}_I)=\sum_{i=1}^{n}R_i(t)\frac{\xi_i(t)\exp\left(\widehat{\gamma}_I I_i(t)\right)}{\sum_{j=1}^{n}\xi_j(t)\exp\left(\widehat{\gamma}_I I_j(t)\right)}.\qquad(7)$$

The estimator $\widehat{\beta}_{BL}$ is, however, biased for the marginal effect of the intervention $I_i(t)$ in our setting of interest, because it is limited to randomized controlled settings and does not consider imbalances between treatment groups which are due to confounders $\mathbf{K_i}(\mathbf{t})$. With a conditional expectation such as in (O2), the process used to build the estimator (which was based on assumption (O1)) is no longer zero-mean and the estimator may thus not converge to the true parameter.

2.4 *The Inverse Probability of Centered Treatment and Monitoring Estimator* $\widehat{\beta}_{IPCTM}$

Under similar assumptions to Bůžková and Lumley (2009), but now further including covariates as in (O2), we first develop an estimator for the conditional effect of $I_i(t)$ on $Y_i(t)$, as in the setting depicted in Figure 1(A). Note that this estimator is marginalized over the predictors $\mathbf{V(t)}$ of the monitoring times and, as in Bůžková and Lumley, we use a monitoring weight to account for any imbalance in those predictors that could bias the effect of $\mathbf{I(t)}$ conditional on $\mathbf{K(t)}$. We define a new process $P_i(t) = P_i(t; \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathscr{A})$ as

$$P_i(t) = \int_0^t \frac{1}{\rho_i(s; \boldsymbol{\gamma})} \left[\left(Y_i(s) - \beta_I I_i(s) - \boldsymbol{\beta_K}' \mathbf{K_i(s)} \right) dN_i(s) - \xi_i(s) \exp\left(\boldsymbol{\gamma_V}' \mathbf{V_i(s)} \right) d\mathscr{A}(s) \right],$$

with $\mathscr{A}(t) = \int_0^t \alpha(s) d\Lambda(s)$. In Web Appendix A, we show that $E\left[dP_i(t) | I_i(t), \mathbf{K_i(t)}\right] = 0$, and the derivation of the estimating equations and estimators for the conditional effects. We obtain the following estimators for the conditional effects of $\begin{bmatrix} I_i(t) & \mathbf{K_i(t)} \end{bmatrix}'$ in (O2):

$$[\widehat{\beta}_I \, \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}]' = \left[\sum_{i=1}^n \int_0^\tau \frac{W(t)}{\rho_i(t; \widehat{\boldsymbol{\gamma}})} \begin{pmatrix} I_i(t) - \overline{I}(t; \widehat{\gamma}_I) \\ \mathbf{K_i(t)} - \overline{\mathbf{K}}(\mathbf{t}; \widehat{\gamma}_\mathbf{I}) \end{pmatrix}^{\otimes 2} dN_i(t) \right]^{-1}$$
$$\times \sum_{i=1}^n \int_0^\tau \frac{W(t)}{\rho_i(t; \widehat{\boldsymbol{\gamma}})} \begin{pmatrix} I_i(t) - \overline{I}(t; \widehat{\gamma}_I) \\ \mathbf{K_i(t)} - \overline{\mathbf{K}}(\mathbf{t}; \widehat{\gamma}_\mathbf{I}) \end{pmatrix}' \left(Y_i(t) - \overline{Y}^*(t; \widehat{\gamma}_I) \right) dN_i(t). \quad (8)$$

Using the estimating equation for conditional effects to estimate the parameters β_I and $\boldsymbol{\beta_K}$ in (O2) corresponds to using a weighted least squares regression with predictors $(I_i(t) - \overline{I}(t; \widehat{\gamma}_I))$ and $(\mathbf{K_i(t)} - \overline{\mathbf{K}}(\mathbf{t}; \widehat{\gamma}_\mathbf{I}))$, a dependent variable $(Y_i(t) - \overline{Y}^*(t; \widehat{\gamma}_I))$ and weights $W(t)/\rho_i(t; \widehat{\boldsymbol{\gamma}})$. To rather estimate the marginal effect of $I_i(t)$ on the mean outcome, we propose to use weights to create a pseudo-population in which there is no imbalance due to confounders, and so we change focus to the corresponding estimating equation for the marginal model (O1), and its corresponding estimator given in (6), when there is no imbalance due to confounders.

The re-weighting procedure we use is reminiscent of standard inverse probability of treatment weighting. Our goal is to break any dependency between the columns of the design matrix in (8), given by $I_i(t)-\overline{I}(t;\widehat{\gamma}_I)$ and $\mathbf{K_i}(\mathbf{t})-\overline{\mathbf{K}}(\mathbf{t};\widehat{\gamma}_\mathbf{I})$. Note that the quantity $\left(I_i(t)-\overline{I}(t;\widehat{\gamma}_I)\right)$ is typically *not* binary so we cannot use a logistic regression to model $E\left[I_i(t)-\overline{I}(t;\widehat{\gamma}_I)|\mathbf{K_i}(\mathbf{t})-\overline{\mathbf{K}}(\mathbf{t};\widehat{\gamma}_\mathbf{I})\right]$. We model the conditional mean using a linear model. Suppose

$$E\left[I_i(t)-\overline{I}(t;\widehat{\gamma}_I)|\mathbf{K_i}(\mathbf{t})-\overline{\mathbf{K}}(\mathbf{t};\widehat{\gamma}_\mathbf{I})\right]=\psi_0+\boldsymbol{\psi_1}'(\mathbf{K_i}(\mathbf{t})-\overline{\mathbf{K}}(\mathbf{t};\widehat{\gamma}_\mathbf{I})). \tag{9}$$

Estimates for $E\left[I_i(t)-\overline{I}(t;\widehat{\gamma}_I)|\mathbf{K_i}(\mathbf{t})-\overline{\mathbf{K}}(\mathbf{t};\widehat{\gamma}_\mathbf{I})\right]$ are obtained via the predictions from the linear regression model (9) with estimated coefficients. To transform these values into pseudo probabilities that lie between 0 and 1 so as to further re-weight the marginal estimating equation corresponding to the estimator in (6), we use an approach suggested by Robins et al. (2000a). We then stabilize these pseudo probabilities, using a marginal model for the mean of $I_i(t)-\overline{I}(t;\widehat{\gamma}_I)$ that is equal to ψ_m so as to compute a final *stabilized generalized weight* given by

$$\text{sgw}_i(t;\widehat{\boldsymbol{\psi}})=\text{sgw}_i(t;\widehat{\psi}_0,\widehat{\boldsymbol{\psi}}_\mathbf{1},\widehat{\psi}_m)=\frac{g^{-1}\left(\widehat{\psi}_0+\widehat{\boldsymbol{\psi}}_\mathbf{1}{}'(\mathbf{K_i}(\mathbf{t})-\overline{\mathbf{K}}(\mathbf{t};\widehat{\gamma}_\mathbf{I}))\right)}{g^{-1}\left(\widehat{\psi}_m\right)} \tag{10}$$

for $g^{-1}(\widehat{a}_i(t))=1/\sqrt{2\pi\widehat{\sigma}_a^2}\exp\left(-\widehat{\epsilon}_{a,i}(t)^2/(2\widehat{\sigma}_a^2)\right)$ the Normal density function evaluated at the linear regression residuals $\widehat{\epsilon}_{a,i}(t)=\left(I_i(t)-\overline{I}(t;\widehat{\gamma}_I)-\widehat{a}_i(t)\right)$, with $\widehat{\sigma}_a^2$ the empirical variance of $\widehat{\epsilon}_{a,i}(t)$. Another way of modelling the variable $I_i(t)-\overline{I}(t;\widehat{\gamma}_I)$ would be to categorize it into quantiles (Naimi et al., 2014). That procedure could work particularly well if the distribution of $I_i(t)-\overline{I}(t;\widehat{\gamma}_I)$ is not unimodal and is asymmetric. This latter approach was evaluated in sensitivity analyses.

The weight (10) is incorporated into the estimating equations corresponding to the estimator of Bůžková and Lumley in (6), and we obtain the new estimating equation

$$\begin{aligned}U^{mar}(\beta,\alpha,\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\psi}})=&\sum_{i=1}^{n}\int_0^\tau\frac{W(t)}{\rho_i(t;\widehat{\boldsymbol{\gamma}})}\frac{1}{\text{sgw}_i(t;\widehat{\boldsymbol{\psi}})}\left(I_i(t)-\overline{I}(t;\widehat{\gamma}_I)\right)\\&\times\left[Y_i(t)-\overline{Y}^*(t;\widehat{\gamma}_I)-\beta\left(I_i(t)-\overline{I}(t;\widehat{\gamma}_I)\right)\right]dN_i(t).\end{aligned} \tag{E2}$$

Solving equation (E2) for $U^{mar}(\beta,\alpha,\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\psi}})=0$ leads to the closed-form solution of our proposed

Inverse Probability of Centered Treatment and Monitoring (IPCTM) estimator, that is given by:

$$\widehat{\beta}_{IPCTM} = \left[\sum_{i=1}^{n}\int_0^{\tau} \frac{W(t)}{\rho_i(t;\widehat{\boldsymbol{\gamma}})} \frac{\left(I_i(t) - \overline{I}(t;\widehat{\gamma}_I)\right)^2}{\text{sgw}_i(t;\widehat{\boldsymbol{\psi}})} dN_i(t)\right]^{-1} \times \sum_{i=1}^{n}\int_0^{\tau} \frac{W(t)}{\rho_i(t;\widehat{\boldsymbol{\gamma}})} \frac{\left(I_i(t) - \overline{I}(t;\widehat{\gamma}_I)\right)}{\text{sgw}_i(t;\widehat{\boldsymbol{\psi}})} \left(Y_i(t) - \overline{Y}^*(t;\widehat{\gamma}_I)\right) dN_i(t) \quad (11)$$

for the estimand of interest, the marginal effect of $I_i(t)$ on $Y_i(t)$.

Note that the intercept function $\alpha(t)$ is left unspecified in (O1) so that one need not assume any particular form for the dependence of the outcome $Y(t)$ on time t. More details on the unbiasedness of the IPCTM estimator are presented in Web Appendix B. Similarly to Bůžková and Lumley (2009), the asymptotic variance of the IPCTM estimator can be developed using standard asymptotic theory. It is also possible to directly account for the components of variance due to the weights using theory on two-step estimators (Newey and McFadden, 1994) along with the variance formula provided by Bůžková and Lumley (2009).

2.5 *The Flexible Inverse Probability of Treatment and Monitoring Estimator* $\widehat{\beta}_{FIPTM}$

A second estimator, which is also a weighted least squares type estimator, is proposed to estimate the marginal effect of treatment on a longitudinal and continuous outcome. It requires slightly stronger parametric specifications for the intercept $\alpha(t)$ in (O1), which is modelled through cubic splines. However, it is easier to implement in practice, and as we will demonstrate in Section 3, it often provides equivalent performance as the IPCTM estimator in simulation studies. Given its more parametric nature, we also expect it to exhibit smaller variance than the IPCTM estimator.

Let us assume again the conditional mean model (O2) along with assumptions (P1), (P2) and (P3) and that monitoring times can be modelled through a proportional intensity model as in (V1). We use a weighted least squares method, and aim to create a pseudo-population in which imbalances due to confounders and covariate-dependent monitoring times are eliminated through re-weighting. We first readjust the observations for the monitoring process using an inverse probability of monitoring weight defined by the inverse of $\varphi_i(t;\boldsymbol{\gamma}_V)$, with

$$\varphi_i(t;\boldsymbol{\gamma}_V) = \exp\left(\boldsymbol{\gamma_1'} \boldsymbol{Z_i(t)} + \gamma_2 I_i(t)\right). \tag{12}$$

Again, assuming a proportional intensity model for the monitoring times, one does not need to estimate the function $\Lambda_0(t)$ in (V1) since this term at time t will cancel out across individuals. The parameters $\boldsymbol{\gamma_1}$ and γ_2 can be estimated by fitting a proportional intensity model.

We use a standard approach to adjust for imbalances due to confounders, and add an inverse probability of treatment weight into the weighted least squares regression. That weight is given by:

$$\mathrm{e}_i(t;\boldsymbol{\omega}) = \frac{1}{\mathbb{I}_{(I_i(t)=1)} P(I_i(t)=1|\mathbf{K_i(t)};\boldsymbol{\omega}) + \mathbb{I}_{(I_i(t)=0)}(1 - P(I_i(t)=1|\mathbf{K_i(t)};\boldsymbol{\omega}))}. \tag{13}$$

The quantities $P(I_i(t)=1|\mathbf{K_i(t)};\boldsymbol{\omega})$ and $P(I_i(t)=0|\mathbf{K_i(t)};\boldsymbol{\omega})$ in (13) can be estimated via logistic regression with $\mathbf{K_i(t)}$ as covariates and $I_i(t)$ as the dependent variable. Once again, knowledge about the problem under study should inform selection of $\mathbf{K_i(t)}$ for inclusion in the treatment model used to estimate the IPT weights in (13).

The intercept $\alpha(t)$ in (O2) is modelled using cubic splines along with a constant intercept. We use splines with two knots and choose the tertiles of the distribution of t for the knots. The final estimator has a closed-form solution given by

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{FIPTM}} = \left[\sum_{i=1}^{n} \int_0^{\tau} \frac{\mathrm{e}_i(t;\boldsymbol{\omega})}{\varphi_i(t;\widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}})} \mathbf{S_i(t)}^{\otimes 2} dN_i(t)\right]^{-1} \sum_{i=1}^{n} \int_0^{\tau} \frac{\mathrm{e}_i(t;\boldsymbol{\omega})}{\varphi_i(t;\widehat{\boldsymbol{\gamma}}_{\boldsymbol{V}})} \mathbf{S_i(t)}' Y_i(t) dN_i(t) \tag{14}$$

with $\mathbf{S(t)}$ a matrix with $s+2$ columns, for s the number of columns in the basis of the cubic spline. The leading column of $\mathbf{S(t)}$ is a vector of 1 for the constant intercept, and the last column corresponds to the intervention $\mathbf{I(t)}$. We are interested in the last coefficient of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{FIPTM}}$, which corresponds to the estimator for the marginal effect of treatment, that we further refer to as $\widehat{\beta}_{FIPTM}$.

The asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{FIPTM}}$ is computed using standard theory on weighted least squares estimator, with the components of variance due to the weights incorporated into the sandwich estimator using theory on two-step estimators (Newey and McFadden, 1994). For derivations, see Web Appendix C. A comparison of the empirical, the bootstrapped and the asymptotic variances in simulation studies is presented in Web Table 3 of Web Appendix D, along with the coverage of the FIPTM estimator.

3. Simulation study

Simulation studies were conducted to assess the performance of both estimators $\widehat{\beta}_{IPCTM}$ and $\widehat{\beta}_{FIPTM}$ for the marginal effect of $I_i(t)$ on the mean of $Y_i(t)$, for different levels of dependency ($\gamma_{\boldsymbol{V}}$) between covariates and monitoring times and for different forms of intercept $\alpha(t)$. The data generating mechanism was similar to the one presented in Figure 1(A) and inspired by Bůžková and Lumley (2009), but incorporates (possibly time-varying) confounders. In a first study described below, the intervention and the confounders were kept time fixed. In a second study, they could vary in time (details are presented in Web Appendix E).

Simulation study 1: Time-fixed confounders and treatment

For all patients i, three baseline confounders $\{K_{1i}, K_{2i}, K_{3i}\}$ were generated with $K_{1i} \sim \text{N}(1, 1)$, $K_{2i} \sim \text{Bernoulli}(0.55)$, and $K_{3i} \sim \text{N}(0, 1)$. The intervention $I_i(t)$ was binary and time-fixed, and was simulated as $I_i \sim \text{Bernoulli}(p_{Ii})$ with $p_{Ii} = \text{expit}\,(0.5 + 0.8K_{1i} + 0.05K_{2i} - 1K_{3i})$. One time-varying mediator $Z_i(t)$ was generated, conditional on I_i, as $Z_i(t)|I_i = 1 \sim \text{N}(2, 1)$ and $Z_i(t)|I_i = 0 \sim \text{N}(4, 4)$. The outcome $Y_i(t)$ was simulated as $Y_i(t) = \alpha(t) + 1\,I_i + 3\,[Z_i(t) - E\,[Z_i(t)|I_i]] + 0.4\,K_{1i} + 0.05\,K_{2i} - 0.6\,K_{3i} + \epsilon_i(t)$ with $\epsilon_i(t) \sim \text{N}(\phi_i, 0.01)$ and $\phi_i \sim \text{N}(0, 0.04)$.

The quantities above were first simulated in continuous time, with time discretized over a grid of 0.01 units, from 0 to τ. Then, monitoring times (i.e. when the outcome was observed) were simulated according to a nonhomogeneous Poisson process, with intensity at time t equal to $\lambda_i(t|I_i, Z_i(t)) = \eta_i \exp\,(\gamma_1 I_i + \gamma_2 Z_i(t))$, with η_i a gamma distributed random variable with mean 1 and variance 0.01. Bernoulli draws with probabilities proportional to these intensities could be used at each time point to assign monitoring times. Monitoring times could be drawn up until the maximum follow-up time τ; we fixed $\tau = 2$ and obtained different mean numbers of visits which depended on parameters (γ_1, γ_2). We tested three combinations: $(\gamma_1, \gamma_2) = (0, 0)$, which corresponded to no dependency on covariates; $(\gamma_1, \gamma_2) = (-0.3, 0.2)$; and $(\gamma_1, \gamma_2) = (0.6, 0.3)$. The follow-up time was further censored at time C_i for each individual, with $C_i \sim$ Uniform$(\tau/2, \tau)$. For $\alpha(t)$, five different functions of time were tested: $\alpha(t) = 3$; $\alpha(t) = 2.5t$; $\alpha(t) = \sin(t)$; $\alpha(t) = \exp(t)$; and $\alpha(t) = \exp(2|\sin(3t)|)$. Two sample sizes, respectively $n = 250$ and $n = 500$, were tested. We used a total of 1000 simulations in each study.

The proposed estimators were compared to more standard ones, i.e. an OLS estimator, a visit-weighted estimator and an IPT-weighted estimator. The OLS estimator $\widehat{\beta}_{OLS}$ was obtained by fitting a linear regression model with outcome $Y_i(t)$, a constant intercept and the independent variable I_i. The estimator $\widehat{\beta}_{VW}$ was a weighted least squares estimator in which a time-dependent monitoring weight was incorporated. The monitoring time model was correctly specified and included I_i and $Z_i(t)$ as explanatory variables. The IPT-weighted estimator was a weighted linear regression estimator in which an inverse probability of treatment weight was incorporated. For the estimators $\widehat{\beta}_{IPCTM}$ and $\widehat{\beta}_{FIPTM}$, the treatment and the monitoring models were correctly specified.

In Web Appendix D, we present the results for 9 additional simulation scenarios in which treatment and confounding variables were also time fixed. Scenarios i) and ii) respectively correspond to the cases where confounder variables $\{K_{1i}, K_{2i}, K_{3i}\}$ were correlated, or where confounder variables $\{K_{1i}, K_{2i}, K_{3i}\}$ affected the monitoring intensity. Scenarios iii) and iv) correspond to the cases where generalized IPT weights in the IPCTM estimator were computed from a cumulative logistic regression, with the variable $I_i(t) - \overline{I}(t; \widehat{\gamma}_I)$ binned into 10 quantiles, or with 20 quantiles, respectively. Sensitivity analyses v), vi), vii) and viii) aim to assess sensitivity to model misspecification via studies where we: v) changed the error distribution for a Log-Normal distribution centered in 0, in the mean outcome model, rather than the Normal errors we previously simulated; vi) incorporated non-linear functions of the confounder covariates in the generative outcome model; vii) incorporated non-linear terms of the covariates in the generative proportional intensity model for the visits; and viii) drew, for each individual, a different intercept function $d\Delta_0(t)$ from 3 possible functions: $d\Delta_0(t) \in \{1; 1.5t; \sin(t)\}$, with respective probability $1/2, 1/4, 1/4$. Finally, the simulation scenario ix) explored the effect of conditioning on confounders in the outcome mean model, for all the estimators that were being compared.

Results

Summary statistics (including empirical bias) for each estimator are found in Web Appendix D. Figure 2 shows absolute biases and empirical mean squared errors (MSE) for each of the five estimators we compared; each boxplot summarizes the distribution of bias or MSE, over all 15 scenarios of dependency and intercept functions that we considered. We also present results for one of the scenarios where $\alpha_0(t) = 3$, in Table 1 in this manuscript. The results in Table 1 were based on a simulation study where exposure and confounders were kept as time-fixed.

[Table 1 about here.]

As we notice in Figure 2, the OLS estimator, which we can see is biased, generally provides variable MSEs due to the different sets of $\boldsymbol{\gamma_V}$ parameters. When adjusting for the monitoring process only, we observe that $\widehat{\beta}_{VW}$ varies much less. However, it remains biased due to confounding. The IPT estimator, on the other hand, is only unbiased when there is no informative visit process. Most importantly, $\widehat{\beta}_{IPCTM}$ and $\widehat{\beta}_{FIPTM}$ exhibit almost zero bias and a quite narrow distribution for their absolute bias. As expected, different parameters (γ_1, γ_2) lead to different mean numbers of visits. Typically, the greater the mean number of visits, the smaller the bias for the two latter estimators (see Web Tables 1 and 2 in Web Appendix D). In Table 1 of this manuscript, we find simular results which are representative of the results from across scenarios. In particular, we find that the absolute bias of the two proposed estimators $\widehat{\beta}_{IPCTM}$ and $\widehat{\beta}_{FIPTM}$ is near 0, but that their variance tends to be greater than that of their comparators, as the $\boldsymbol{\gamma}_V$ coefficients increase.

The two proposed estimators dramatically outperform their comparators in terms of bias as those coefficients increase.

In Figure 2, we also observe that the IPCTM estimator exhibits a greater MSE than the flexible estimator ($\widehat{\beta}_{FIPTM}$) in studies with time-fixed treatment and confounder variables, while it exhibits a smaller mean squared error than the FIPTM estimator in studies with time-varying treatment and confounder variables. As expected, the range of MSE narrows as the sample size increases. Given that both $\widehat{\beta}_{IPCTM}$ and $\widehat{\beta}_{FIPTM}$ exhibit a bias that tends towards 0, and that $\widehat{\beta}_{FIPTM}$ is easier to implement in practice, we contend that it should be preferred. We present in Web Table 3 of Web Appendix D a comparison of its bootstrapped, empirical and asymptotic variances, which were generally very similar. In studies with time-varying treatment and confounding variables, the IPCTM estimator may be more efficient. Further investigation of whether the centered estimator may be more competitive in a wider range of scenarios will be an important avenue of future work.

[Figure 2 about here.]

Sensitivity analyses for the first simulation study with time-fixed treatment and confounders

The results (distributions of biases and MSEs) for all 9 sensitivity analyses can be found in Web Tables 4 (i), 5 (ii), 6 (iii and iv), 7 (v, vi, vii, viii), and 8 (ix) in Web Appendix D. A brief summary of these results can also be found in Web Appendix F. Overall, our proposed methods were not too sensitive to misspecification of the different models, except for the sensitivity analysis where we incorporated non-linear functions of the covariates in the proportional intensity model for monitoring times. In that latter case, the FIPTM estimator has shown great bias, while the IPCTM estimator was not as affected by the misspecification of the monitoring model.

4. Application to the *Add Health* Study

The proposed estimators were applied to data from the National Longitudinal Study of Adolescents to Adult Health (*Add Health*) (Harris et al., 2009a) to assess the marginal effect of depressive mood on weight in pounds, in adolescents. Our estimators were also compared to more standard estimators that do not account for informative monitoring process and/or confounding.

Add Health is a four-wave longitudinal study on adolescents who, over the course of the study, age to become young adults. A pool of participants who were well representative of adolescents in United States were enrolled during the years 1994-5 while they were in grades 7 to 12, and followed until 2008 (Wave IV). For each of the four waves, an in-home questionnaire was completed by the participants. A parent questionnaire was completed by one of the participants' parents at baseline only (Wave I). Data collected from in-home interviews are publicly available online for all four waves (Harris, 2009b). For the purpose of this analysis, we assumed that longitudinal data are made up of a maximum of four time points where the outcome is potentially recorded. Hence, $time = 1, 2, 3, 4$ respectively correspond to all four waves. For simplicity, none of our analyses considered the sampling weights used in *Add Health* study.

We first defined the time-varying exposure that consisted of a binary depression score, using a question from the in-home interview that was related to the current depressive mood of the participant. For the question *How often was the following true during the past week? You felt depressed.*, a participant's score was set to 0 if they answered *Never/rarely* or *Sometimes* and to 1 if their answer was contained in *A lot of the time* or *Most/All of the time*. The longitudinal outcome consisted in the weight in pounds, which was recorded at every in-home interview. We assumed that the relationship between depression status and weight was mediated by smoking, since depressive mood exacerbates smoking (e.g., Stepankova et al., 2016), which in turn affects weight (e.g., Grunberg, 1985). We used as a proxy for smoking the number of cigarettes smoked during the past 30 days, also recorded at each of the four in-home interviews. A participant who had smoked at least one cigarette in the previous 30 days was considered to be a smoker. For confounders of the relationship between depression and weight, we included age, sex and socioeconomic status (SES). SES was defined using the two following in-home questions asked to one of the participants' parents: *About how much total income, before taxes did your family receive in 1994?* and *How far did you go in school?*. The answers were transformed into quintiles and summed up to give a total score contained between 0 and 10, with 10 corresponding to the highest SES.

A total of 6504 participants were enrolled at Wave I. Data presented missing values due to patients' dropout or their refusal to answer questions during the course of the study. We assumed that monitoring times (i.e. times when weight was recorded) depended on the depression status, the smoking status, age, sex and SES, which variables were included in a proportional intensity model for the monitoring times. In the exposure model, we adjusted for the potential confounders age, sex and SES. If patients had a value at their first interview, this value was used to impute values at other waves (it remained fixed in time). Recall that variables predicting the visit process are required to be available at all time. Thus, we employed multiple imputation with M=5 imputations, using predictive mean matching to impute any remaining missing values in age, sex and SES, as well as for missing values in exposure and mediator. Following imputation and analysis, the coefficient for exposure of interest was combined across the imputations (Rubin, 1976). One thousand stratified bootstrap samples were drawn, with strata taken to be the individual, and they were used to assess the variance of each of the 5 estimates we compared. Table 2 presents a summary of the characteristics of the cohort at baseline, stratified by their depressive mood. Table 3 presents the average rate ratios for the 5 variables that were incorporated into the proportional intensity model for the visit times, along with confidence intervals computed using Rubin's rule for multiply imputed datasets (Rubin, 2004). Table 4 shows all estimated effects of depressive mood on weight with corresponding 95% Wald-type confidence intervals (CIs) using bootstrap standard errors.

[Table 2 about here.]

The two exposure groups (depressed/not depressed) presented differences at baseline, with more smokers, older participants, more females and lower SES on average in the participants with depressive mood than in those without. Smoking and sex (female) were associated with a higher probability of the outcome being reported, and age with a lower probability (Table 3).

[Table 3 about here.]

[Table 4 about here.]

An important difference was found between the estimates for the marginal effect of depressive mood computed using $\widehat{\beta}_{OLS}$, $\widehat{\beta}_{VW}$, or $\widehat{\beta}_{IPT}$, and those obtained with our proposed estimators. The change in estimate seemed to be due to both confounding and informative monitoring times, with an important difference between $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_{IPT}$, and an important remaining difference between $\widehat{\beta}_{IPT}$ and $\widehat{\beta}_{FIPTM}$ or $\widehat{\beta}_{IPCTM}$. The methods that did not account for confounding and informative monitoring times suggest that depressive mood leads to decreases in weight of nearly four pounds.

After adjusting for confounding and informative monitoring times, the estimates were consistent with those found in the literature. We found a small increase in weight due to depressive mood, with the lower limit of the confidence interval that corresponded to a weight loss of about half a pound, and an upper limit that consisted of a weight gain of just over three pounds. Wurtman (1993) explained the complex relationships leading to weight increase in patients with depression and the link with smoking. Studies such as Van Strien et al. (2016) found no significant direct effect of depression on weight gain but only a positive effect through emotional eating as a mediator.

The differences observed and the sign reversal of the estimates after accounting for important features that may bias the estimates echo the results of Hernán et al. (2000). The fact that we observed a reversal between the IPT-weighted estimator and the FIPTM and IPCTM estimators supports the message that informative visit process-induced bias should be accounted for.

5. Discussion

Electronic health records are increasingly available and a common source of data to study the effect of treatments on longitudinal outcomes in pharmacoepidemiological studies (Hennessy, 2006). Given their real-world nature, monitoring times in EHRs are often covariate-dependent and the outcome recorded may be associated with the same covariates, which introduces selection bias in the analysis. Most often, that feature is ignored. However, when it is considered, confounding bias is rarely accounted for, as – until now – no simple method has been described to account for the two sources of bias simultaneously. In this article, we proposed two novel estimators for the marginal effect of a treatment on a longitudinal outcome which account for imbalances due to covariate-dependent monitoring times, confounding and mediation. Neither estimator requires the longitudinal outcome to be measured at all times in continuous but rather only sporadically. The asymptotic properties of both estimators can be derived. These estimators are relevant to EHRs and to studies where irregular monitoring times were planned.

The proposed estimators were compared to more standard ones in simulation studies and both outperformed the OLS estimator, the weighted least squares estimator with an inverse monitoring weight and the inverse probability of treatment weighted estimator. Their empirical absolute bias tended towards 0, and the FIPTM estimator has shown good coverage. Moreover, we provided a practical framework for analysts, with both estimators being flexible with regards to the modelling of the intercept function. We recommend the use of the FIPTM estimator, which is easy to implement in practice and for which we have derived the asymptotic variance. For situations where the intercept function $\alpha(t)$ is expected to vary extensively in time, or for time-dependent treatment and confounders settings, the IPCTM estimator could be preferred and has shown to be well-behaved.

The estimators we propose rely on important assumptions. One challenge related to this work is the need for the treatment model to be correctly specified, and the risk for unmeasured confounding. Unmeasured confounding has been widely discussed, and sensitivity analyses are available to evaluate the degree at which it could impact the estimate of interest (Lash and Fink, 2003; Schneeweiss, 2006; Robins, Rotnitzky, and Scharfstein, 2000b). In the situation where the treatment model is misspecified, the IPT weights may not provide adequate adjustment for confounding. Knowledge about the research problem should inform the set of potential confounders to incorporate into the treatment model. The use of directed acyclic graphs may help in determining which predictors should be included in the treatment model (Pearl, 1995; Greenland et al., 1999), however these encode the analyst's beliefs and may themselves overlook important variables.

Another challenge is the need for the predictors of the monitoring process to be recorded at all times. In administrative databases and EHRs, information on drugs prescribed or dispensed, diagnostics and interventions are often recorded even in between physician visits when the outcome is monitored. For instance, in a study where the question is whether a particular drug impacts the outcome of blood pressure, blood pressure might be measured only when a patient's physician suspects changes in blood pressure and yet the patient potentially visited the physician at several other points, with data such as the exposure and comorbidities being recorded. In some observational studies, however, it will not be possible to assess covariate values in between the times when the outcome is measured. In that case, our methods could be extended to incorporate only the covariates measured at monitoring times, and to use them to predict the future monitoring times.

Acknowledgements

We would like to thank Professor Thomas Lumley for his generous assistance with the application of methods proposed in *Semiparametric modeling of repeated measurements under outcome-dependent follow-up* (Bůžková and Lumley, 2009). We are also very grateful to the Editor, the Associate Editor and two referees who provided helpful comments that greatly improved the manuscript.

This work is supported by a doctoral scholarship from the Natural Sciences and Engineering

Research Council (NSERC) of Canada (Ref. 401223940) to author JC. EEMM acknowledges support from a Discovery Grant from NSERC and a chercheur-boursier career award from the Fonds de recherche du Québec – Santé.

This research uses data from *Add Health*, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. No direct support was received from grant P01-HD31921 for this analysis.

Data availability statement

The data that support the findings in this paper are openly available in the Data Sharing for Demographic Research repository, at https://doi.org/10.3886/ICPSR21600.v21. The analysis was restricted to the Add Health public-use data and did not include restricted-use data (Harris and Udry, 2018).

References

Bůžková, P., and Lumley, T. (2005). Marginal regression modeling under irregular, biased sampling. *UW Biostatistics Working Paper Series*. Working paper 261.

Bůžková, P., and Lumley, T. (2009). Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Statistics in Medicine*, 28(6), 987–1003.

Cai, N., Lu, W., and Zhang, H. H. (2012). Time-varying latent effect model for longitudinal data with informative observation times. *Biometrics*, 68(4), 1093–1102.

Cao, H., Li, J., Fine, J. P., *et al.* (2016). On last observation carried forward and asynchronous longitudinal regression analysis. *Electronic Journal of Statistics*, 10(1), 1155–1180.

Greenland, S., and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3), 413–419.

Greenland, S., Pearl, J., Robins, J. M., *et al.* (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48.

Grunberg, N. E. (1985). Nicotine, cigarette smoking, and body weight. *British Journal of Addiction*, 80(4), 369–377.

Harris, K. M. (2009). The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007-2009 [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Harris, K. M., Halpern, C. T., *et al.* (2009). The National Longitudinal Study of Adolescent to Adult Health: Research Design document *The Add Health Study: Design and Accomplishments*. Chapel Hill, NC: Carolina Population Center, University of North Carolina-Chapel Hill. https://www.cpc.unc.edu/projects/addhealth/documentation/guides/DesignPaperWIIV.pdf (accessed January 28, 2019).

Harris, K. M., and Udry, J. R. (2018). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR21600.v21 (accessed May 6, 2018).

Hennessy, S. (2006). Use of health care databases in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, 98(3), 311–313.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5), 561–570.

Lash, T. L., and Fink, A. K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*, 14(4), 451–458.

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 44(4), 121–130.

Lin, D. Y., and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 96(453), 103–126.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., *et al.* (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, 58(3), 621–630.

Naimi, A. I., Moodie, E. EM, Auger, N., *et al.* (2014). Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*, 25(2), 292–299.

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4, 2111–2245.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments – essay on principles. *Annals of Agricultural Sciences*, section 9, 10, 1–51.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688.

Pullenayegum, E. M., and Lim, L. S. H. (2016). Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, 25(6), 2992–3014.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429), 106–121.

Robins, J. M., Hernán, M. A., and Brumback, B. A. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 116, 1–94.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rosenbaum, P. R. (1985). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A*, 147(5), 656–666.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Ryu, D., Sinha, D., Mallick, B., *et al.* (2007). Longitudinal studies with outcome-dependent follow-up. *Journal of the American Statistical Association*, 102(479), 952-–961.

Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*, 15(5), 291–303.

Stepankova, L., Kralikova, E., Zvolska, K., *et al.* (2016). Depression and smoking cessation: evidence from a smoking cessation clinic with 1-year follow-up. *Annals of Behavioral Medicine*, 51(3), 454–463.

Sun, J., Park, D.-H., Sun, L., *et al.* (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*, 100(471), 882–889.

Sun, L., Song, X., Zhou, J., *et al.* (2012). Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *Journal of the American Statistical Association*, 107(498), 688–700.

Tan, K. S., French, B., and Troxel, A. B. (2014). Regression modeling of longitudinal data with outcome-dependent observation times: extensions and comparative evaluation. *Statistics in Medicine*, 33(27), 4770–4789.

Van Strien, T., Konttinen, H., Homberg, J. R., *et al.* (2016). Emotional eating as a mediator between depression and weight gain. *Appetite*, 100, 216–224.

Wurtman, J. J. (1993). Depression and weight gain: the serotonin connection. *Journal of Affective Disorders*, 29(2-3), 183–192.

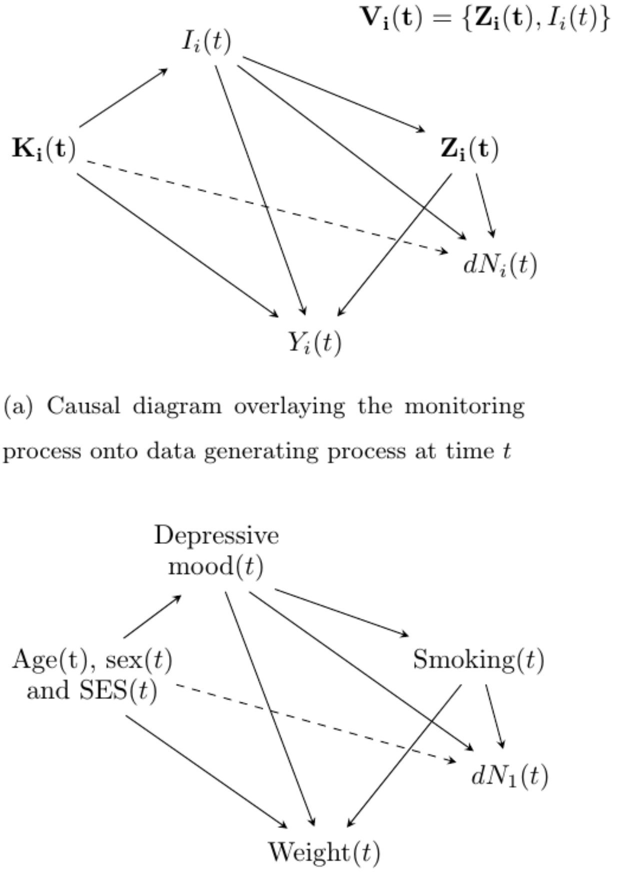
Zhu, Y., Lawless, J. F., and Cotton, C. A. (2017). Estimation of parametric failure time distributions based on interval-censored data with irregular dependent follow-up. *Statistics in Medicine*, 36(10), 1548–1567.

Supporting Information

Web Appendices A, B, C, D, E and F referenced in Sections 2 and 3, and the R code to reproduce the simulations (for time-fixed treatment and confounders scenario) and to compute the proposed estimators are available with this paper at the Biometrics website on Wiley Online Library.

Received July 2019. *Revised March* 2020. *Accepted April* 2020.

LIST OF FIGURES



(a) Causal diagram overlaying the monitoring process onto data generating process at time t

(b) Causal diagram for the *Add Health* study data at time t in individual $i = 1$

Figure 1: Structure of the data generating and monitoring process for (a) a general setting and (b) the analysis of *Add Health* data

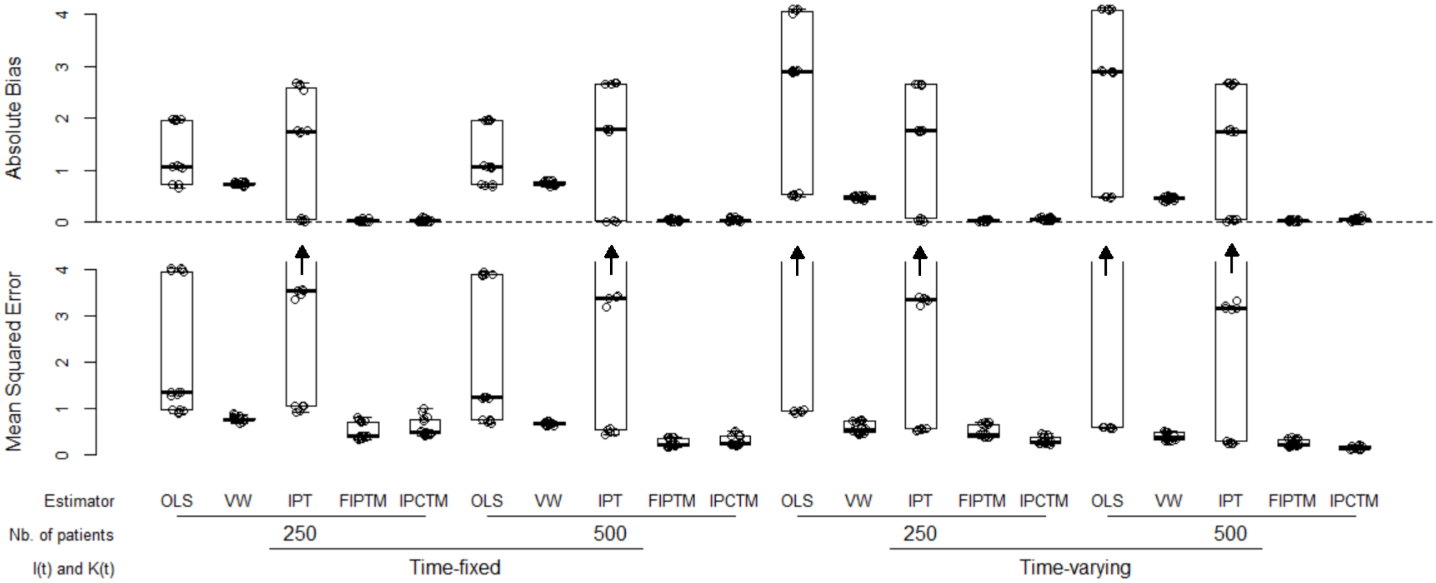


Figure 2: Boxplots of the distribution of absolute bias (top panel) and of MSE (bottom panel) from all 15 simulation scenarios, for the five estimators: Ordinary least squares, visit weighted only, inverse probability of treatment weighted estimator, FIPTM and IPCTM estimator, for time-fixed (left) or time-varying variables (right) and different sample sizes ($\tau = 2$, 1000 simulations).

LIST OF TABLES

Table 1: Simulation study with confounding and covariate-dependent monitoring times ($\tau = 2$, $n = 250$, $\alpha(t) = 3$, time-fixed exposure and confounders)

(γ_1,γ_2)	Median no.	Absolute bias (Empirical variance)				
	visits (IQR)	$\hat{\beta}_{OLS}$†	$\hat{\beta}_{VW}$‡	$\hat{\beta}_{IPT}$⋆	$\hat{\beta}_{FIPTM}$	$\hat{\beta}_{IPCTM}$
$(0, 0)$	1 (1-2)	0.72 (0.41)	0.71 (0.30)	0.06 (1.06)	0.09 (0.77)	0.08 (0.99)
$(-0.3, 0.2)$	2 (1-3)	1.05 (0.19)	0.72 (0.18)	1.77 (0.40)	0.04 (0.39)	0.01 (0.44)
$(0.6, 0.3)$	5 (4-7)	1.98 (0.12)	0.76 (0.19)	2.65 (0.30)	0.00 (0.38)	0.02 (0.47)

† Ordinary least squares regression with outcome $Y_i(t)$ and exposure $I_i(t)$ with a constant intercept

‡ Weighted least squares regression with outcome $Y_i(t)$ and exposure $I_i(t)$ with a constant intercept and an inverse probability of monitoring weight computed from a proportional intensity model with $I_i(t)$ and $Z_i(t)$ as predictors

⋆ Weighted least squares regression with outcome $Y_i(t)$ and exposure $I_i(t)$ with a constant intercept and one an inverse probability of treatment weight computed from a logistic regression model with $\mathbf{K_i(t)}$ as predictors

Table 2: Characteristics at baseline of children enrolled in the *Add Health* study, stratified by depressive mood

	Depressive mood	
Variable	No	Yes
Smoking (N, %)	1367 (23.3)	280 (44.0)
Age (median, IQR)	15 (14-16)	16 (14-17)
Sex=female (N, %)	2914 (49.8)	433 (68.0)
SES (median, IQR)	6 (4-8)	5 (4-7)

Table 3: Average rate ratios and 95% confidence intervals for variables in the proportional intensity model for monitoring times

Variable	Rate ratio	95 % CI
Depressive mood	0.93	0.84; 1.02
Smoking	1.08	1.03; 1.13
Age	0.94	0.93; 0.94
Sex=female	1.04	1.01; 1.07
SES	1.00	0.99; 1.01

Table 4: Comparison of the estimates of the marginal effect of depression status on average weight in pounds

	Estimate	95% CI
$\widehat{\beta}_{OLS}$	-3.83	-5.55; -2.11
$\widehat{\beta}_{VW}$	-3.69	-5.44; -1.94
$\widehat{\beta}_{IPT}$	-1.56	-3.45; 0.33
$\widehat{\beta}_{FIPTM}$	1.43	-0.35; 3.21
$\widehat{\beta}_{IPCTM}$	1.12	-0.59; 2.83