Received: 7 April 2022 | Revised: 10 February 2023 | Accepted: 18 March 2023

DOI: 10.1002/sim.9727

RESEARCH ARTICLE

Statistics in Medicine WILEY

# Causal inference with longitudinal data subject to irregular assessment times

**Eleanor M. Pullenayegum**[1,2] | **Catherine Birken**[1,3] | **Jonathon Maguire**[4,5,6] | **The TARGet Kids! Collaboration**[1,6]

[1]Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, Canada

[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

[3]Department of Paediatrics, University of Toronto, Toronto, Canada

[4]Department of Paediatrics, St Michael's Hospital, Toronto, Canada

[5]Departments of Paediatrics & Nutritional Sciences, University of Toronto, Toronto, Canada

[6]Li Ka Shing Knowledge Institute, Unity Health Toronto, Toronto, Canada

**Correspondence**
Eleanor M. Pullenayegum, Child Health Evaluative Sciences, Hospital for Sick Children, 555 University Ave, Toronto, ON, M5G 1X8, Canada.
Email: eleanor.pullenayegum@sickkids.ca

**Funding information**
Natural Sciences and Engineering Research Council of Canada

**Summary**

Data collected in the context of usual care present a rich source of longitudinal data for research, but often require analyses that simultaneously enable causal inferences from observational data while handling irregular and informative assessment times. An inverse-weighting approach to this was recently proposed, and handles the case where the assessment times are at random (ie, conditionally independent of the outcome process given the observed history). In this paper, we extend the inverse-weighting approach to handle a special case of assessment not at random, where assessment and outcome processes are conditionally independent given past observed covariates and random effects. We use multiple outputation to accomplish the same purpose as inverse-weighting, and apply it to the Liang semi-parametric joint model. Moreover, we develop an alternative joint model that does not require covariates for the outcome model to be known at times where there is no assessment of the outcome. We examine the performance of these methods through simulation and illustrate them through a study of the causal effect of wheezing on time spent playing outdoors among children aged 2–9 years and enrolled in the TargetKids! study.

**KEYWORDS**

causal inference, inverse weighting, joint modeling, longitudinal data

## 1 | INTRODUCTION

Longitudinal data collected in the context of usual care are becoming increasingly available for research. Because this data is gathered as part of usual care, it is not optimized for research purposes and consequently poses analytic challenges. We consider two challenges in particular. First, determining the causal effect of an exposure is complicated by the fact that the exposure is usually not randomized and hence is subject to confounding. Second, the times at which outcomes are assessed often differ among patients and may be associated with the outcome of interest: for example, often outcomes are assessed when a patient visits their physician, and patients may be followed more often when unwell. While the issues of causal inference and irregular assessment times have been discussed at length, their co-occurrence has been only

recently considered.[1] The purpose of this paper is to extend these recently developed methods to handle a wider range of informative assessment times.

There are two main classes of methods for handling irregular and informative assessment times: approaches based on inverse intensity weighting,[2,3] and approaches based on joint models, which can be either semi-parametric[4-7] or fully parametric.[8,9] Each class of methods requires a slightly different set of assumptions about the inter-relationships between the outcome and assessment time processes. Inverse intensity weighting assumes that the outcome and assessment time processes are conditionally independent given previously observed outcomes, covariates and assessment times;[2] this is known as assessment at random (AAR).[10] Semi-parametric joint models assume that the outcome and assessment time processes are conditionally independent given baseline covariates and random effects;[8,9] this is a special case of assessment not at random (ANAR).[10] The choice of analytic procedure thus hinges on the assumed dependence between the assessment time and outcome processes.[11]

Approaches for causal inference with longitudinal observational data include marginal structural models (MSMs),[12] g-computation,[13] and targeted maximum likelihood estimation (TMLE).[14] Marginal structural models[12] are a popular approach, and rely on weighting GEE equations by the inverse of the probability of treatment. They assume that the assessment times are not informative about the outcome of interest, that is, that assessment is completely at random (ACAR).[10]

Coulombe et al[1] devised two approaches to allow for causal inference under AAR: multiplying the inverse-intensity weights by the inverse-probability weights and using a GEE (hereafter referred to as doubly-weighted GEE or DW-GEE), or by adapting Buzkova & Lumley's weighted Lin-Ying equations.[15,16] These models inherit the assumptions of inverse-intensity weighting and marginal structural models, and hence are not suitable for data with ANAR.

In practice, we may wish to do inference under special cases of ANAR. There may be factors that predict both outcome and assessment frequency that are not recorded in the data, for example job security. When these factors are stable over time, the relationship can be captured using random effects. We propose to apply Coulombe's double weighting approach to semi-parametric joint models so as to allow for random effects that predict both the outcome and the assessment intensity. To do so, we must overcome the challenge that the estimation procedure for semi-parametric joint models cannot accommodate weights that vary within subjects.[17]

Multiple outputation has been previously proposed as an approach to handling irregular assessment times in settings where weighting cannot be used.[17] Where multiple imputation imputes missing data multiple times, multiple outputation discards excess data multiple times.[18,19] Multiple outputation for longitudinal data subject to irregular assessment times works by selecting observations with probability inversely proportional to the intensity of the assessment time process, thus creating a thinned dataset in which the outcome and assessment time processes are conditionally independent given random effects.[17] While multiple outputation has been shown to handle informative assessment times, it has never been used in the context of causal inference, where there is also a time-varying exposure subject to confounding.

In this paper, we develop the first approach to causal inference in the presence of a time-varying exposure and longitudinal data with irregular assessment times that are not at random. Under the assumption that the outcome and assessment times are independent conditionally on random effects and previously observed data, we propose to use multiple outputation to construct thinned datasets in which the exposure-outcome relationship is unconfounded and in which, conditionally on random effects, the outcome at any time $t$ is independent of the assessment process at time $t$. We show how these thinned datasets can be used with semi-parametric joint models to accommodate correlated random effects in the outcome and assessment time models. We also propose an alternative estimation procedure for semi-parametric joint models that does not require the outcome model covariates to be known at times when an assessment does not occur.

We describe the methods in detail in Section 2, explore their finite sample performance through simulation in Section 3 and use our methods to explore the causal effect of wheezing on time spent playing outdoors in Section 4. We conclude with a discussion in Section 5.

## 2 | CAUSAL INFERENCE WITH IRREGULAR OBSERVATION

### 2.1 | Notation

For a subject $i$ ($i = 1, \ldots, n$), at time $t \in [0, \tau)$, let $Y_i(t)$ be the outcome, which we assume has interval measurement properties, let $A_i(t)$ indicate whether an exposure occurred at time $t$, and let $Z_i(t)$ be the observed covariates, which are auxiliary in the sense that we do not wish to condition our inference about $Y$ on them. We assume that outcomes $Y_i$ are observed

only when a patient visits their physician, but that $Z_i$ and $A_i$ are fully observed. As in Coulombe et al,[1] we consider point exposure effects, such that it is only the current value of $A_i$ that affects outcome, not the whole history. Let $T_{ij}$ be the times at which outcomes $Y_i$ are assessed, with corresponding counting process $N_i$, and let $C_i$ be the censoring time for subject $i$. The observed history for subject $i$ at time $t$ is $\mathcal{H}_i(t) = \{I(C_i > t), C_i I(C_i < t), \overline{Z}_i(t \wedge C_i), \overline{N}_i(t^-), \bar{A}(t \wedge C_i), \Delta N_i(s) Y_i(s) : s < t\}$, where, for an arbitrary process $B$, $\overline{B}(t) = \{B(s) : s \leq t\}$ and $\Delta B(t) = B_i(t) - B_i(t^-)$.

Let $Y_i^a(t)$ be the outcome at time $t$ when subject $i$ experiences the (possibly counterfactual) exposure $A(t) = a$, where for simplcity we assume that $A_i(t) \in \{0, 1\}$. Define $N_i^a(t)$ as the counting process corresponding to the observation times for $Y_i^a$, and note that $N_i(t) = N_i^0(t) + N_i^1(t)$. Let $\eta_{Vi}$ and $\eta_{Yi}$ be potentially correlated random effects for the assessment time and outcome processes with means 1 and 0 respectively. We consider models of the form

$$E(Y_i^a(t)|\eta_{Yi}) = \alpha_0(t) + a\beta_0 + W(a,t)\eta_{Yi} \tag{1}$$

where $\alpha_0(t)$ is a (possibly non-parametric) function of time $t$, and $W$ is a vector of deterministic functions of $a$ and $t$ corresponding to the random effects $\eta_{Yi}$ (eg, an intercept).

## 2.2 | Causal estimand

Our target of inference will be the marginal effect of the exposure on the outcome recorded at the same observation time, that is, $E(Y_i^1(t) - Y_i^0(t)) = \beta_0$. Note from Equation (1) that this is time-invariant. In a population in which exposures and outcomes and not confounded and when assessment times are independent of outcomes, this could simply be estimated by $E(Y_i(T_{ij})|A_i(T_{ij}) = 1) - E(Y_i(T_{ij}) \,|\, A_i(T_{ij}) = 0)$. Since this is not the case in our data we require an alternative analytic approach.

## 2.3 | Assumptions

Figure 1 provides an overview of the assumptions, with time is discretized over a fine grid. We provide detailed assumptions about the treatment and assessment processes below.

### 2.3.1 | Treatment process

We assume that, for some sub-history $\mathcal{H}^*(t^-)$ of the full history $\mathcal{H}(t^-)$,

(a) $1 > P(A_i(t) = a|\mathcal{H}_i^*(t^-)) > 0$ whenever $P(\mathcal{H}_i^*(t^-)) > 0$ (positivity)
(b) $Y_i^a(t) \perp\!\!\!\perp A_i(t)|\mathcal{H}_i^*(t^-)$ (no unmeasured confounders)
(c) $Y_i^a(t)|A_i(t) = Y_i^a(t)|A_i'(t)$ whenever $A_i(t) = A_i'(t)$ (stable-unit treatment value)

### 2.3.2 | Assessment time process

We assume that the assessment time process has intensity

$$\begin{aligned}\lambda_i(t) &= \lim_{\delta \downarrow 0} \frac{E(N_i(t) - N_i(t-\delta)|\mathcal{H}_i^*(t-\delta), \eta_{Vi}, \eta_{Yi}, A_i(t), Z_i(t), Y_i(t))}{\delta} \\ &= \lim_{\delta \downarrow 0} \frac{E(N_i(t) - N_i(t-\delta)|\mathcal{H}_i^*(t-\delta), A_i(t), \eta_{Vi}, Z_i(t))}{\delta}\end{aligned}$$

that is, we assume that the assessment times are conditionally independent of outcomes given the observed history, auxiliary covariates and assessment time random effect. We additionally assume that the treatment assignment process is independent of $\eta_{Vi}$, so that the intensity for the observations times of $Y^a$ can be expressed as

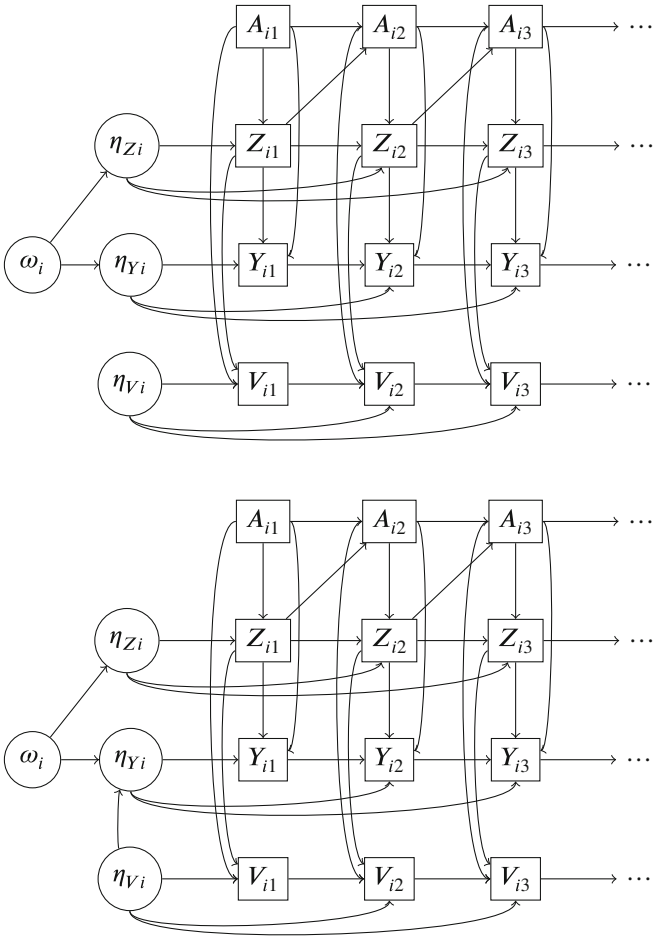


**FIGURE 1** Directed acyclic graphs depicting relationships between outcomes ($Y$), assessment indicators $V$, exposures ($A$), auxiliary covariates ($Z$) and random effects $\eta_{Yi}, \eta_{Zi}, \eta_{Ti}$. For the purposes of illustration, time has been discretized on a fine grid so that $V_{ij}$ is an indicator representing whether there was an outcome assessment for subject $i$ at time $j$, and $Y_{ij}$ is the outcome for subject $i$ at time $j$. Note that $Y_{ij}$ is observed only if $V_{ij} = 1$. Top panel: (A) Outcome and assessment time processes conditionally independent given auxiliary covariates; Bottom panel (B): Outcome and assessment time processes conditionally independent given auxiliary covariates and random effects.

$$
\begin{aligned}
\lambda_i^a(t) &= \lim_{\delta \downarrow 0} \frac{E(N_i^a(t) - N_i^a(t-\delta) | \mathcal{H}_i^*(t-\delta), \eta_{Vi}, \eta_{Yi})}{\delta} \\
&= \lim_{\delta \downarrow 0} \frac{E(N_i(t) - N_i(t-\delta) | \mathcal{H}_i^*(t-\delta), A_i(t) = a, \eta_{Vi}) P(A_i(t) = a | \mathcal{H}_i^*(t-\delta), \eta_{Yi}, \eta_{Vi})}{\delta} \\
&= \lim_{\delta \downarrow 0} \frac{E(N_i(t) - N_i(t-\delta) | \mathcal{H}_i^*(t-\delta), A_i(t) = a, \eta_{Vi}) P(A_i(t) = a | \mathcal{H}_i^*(t-\delta))}{\delta} \quad \text{(no unmeasured confounders)} \\
&= E(\lambda_i(t) | \mathcal{H}_i^*(t^-), \eta_{iV}, A_i(t) = a) P(A_i(t) = a | \mathcal{H}_i^*(t^-)).
\end{aligned}
$$

Note that if the assessment intensity $\lambda$ does not depend on $Z_i(t)$ given the observed history $\mathcal{H}(t^-)$, then we simply have

$$\lambda_i^a(t) = \lambda_i(t) P(A_i(t) = a | \mathcal{H}_i^*(t^-)).$$

More generally, however, $Z_i(t)$ may influence the assessment intensity and also mediate the exposure outcome relationship as in Figure 1, so that the assessment intensity must be conditioned on $Z_i(t)$ while the exposure probability should not be.

We assume that censoring is non-informative in the sense that

$$E(Y_i(t) | A_i(t), C_i) = E(Y_i(t) \,|\, A_i(t)).$$

As can be seen in Figure 1, there is a backdoor path between the indicator $V_{ij}$ for whether an assessment occurred at time $j$ and the outcome $Y_{ij}$ via both the auxiliary covariates $Z_{ij}$ and the exposure $A_{ij}$. Furthermore, $Z$ both confounds the exposure-outcome relationship and is a mediator. Inference about the causal effect of $A_{ij}$ on $Y_{ij}$ must thus account for both the confounding effect of $Z$ and the informative nature of the assessment process. Coulombe et al[1] show how this can be done using a doubly weighted GEE in scenario 1 (a) where the random effects for outcome and assessment time processes are independent. We seek to study scenario 1 (b), where the random effects are dependent.

## 2.4 | Model

We impose a proportional hazards assumption on the assessment intensity for the outcomes, that is:

$$\lambda_i(t) = \lim_{\delta \downarrow 0} \frac{E(N_i(t) - N_i(t-\delta) | \mathcal{H}_i^*(t-\delta), A_i(t) = a, \eta_{Vi}, \eta_{Yi}, Z_i(t))}{\delta} = I(C_i \geq t)\lambda_0(t)\eta_{Vi} \exp(Z_i(t)\gamma_0 + A_i(t)\gamma_{0a}),$$

where for notational convenience we allow $Z_i(t)$ to include past observed outcomes, exposures and observation times. We thus assume that the intensity with which outcomes are assessed follows a proportional hazards assumption for covariates $Z_i(t)$ and exposure $A_i(t)$, with a multiplicative frailty $\eta_{Vi}$. Note that for $\gamma_0$ and $\gamma_{0a}$ in this model to be estimable we require $Z_i(t)$ and $A_i(t)$ to be observed at all times. Writing $\gamma_{0*} = (\gamma_0, \gamma_{0a})$ and $Z_{ai}(t) = (Z_i(t), A_i(t))$ we have

$$\lambda_i(t) = I(C_i \geq t)\lambda_0(t)\eta_{Vi} \exp(Z_{ai}(t)\gamma_{0*})$$

We assume that the treatment assignment probabilities depend on past covariates $Z_i$:

$$P(A_i(t) = a | \mathcal{H}_i^*(t^-)) = \pi\left(a; Z_i(t^-)\right),$$

For the outcome $Y^a$ we use the linear model with random effects in Equation (1), namely:

$$Y_i^a(t) = \alpha_0(t) + a\beta_0 + W(a,t)\eta_{iY} + \epsilon_i(t) \tag{2}$$

where $\alpha_0$ is an unspecified function of time, $\beta_0$ is the parameter of interest and $\epsilon_i(t)$ is a residual with $E(\epsilon_i(t)|A_i(t), \eta_{Yi}, \eta_{Vi}) = 0$. Note that it is straightforward to add fixed effect covariates to Equation (2) provided that they are time-invariant; we have omitted them for simplicity.

In Equation (2), the random effects $\eta_{Yi}$ represent unmeasured baseline covariates that affect the trajectory of the outcome, for example genetic or socioeconomic variables, while the frailty variable $\eta_{Vi}$ represents unmeasured baseline variables that affect the frequency with which patients have their outcomes assessed. These could include risk factors for poor outcomes that are not included in the data, or socioeconomic variables; for example in our previous work we found that higher income and higher levels of education were associated with more frequent primary care visits in a paediatric population.[20] The outcome random effects may thus be correlated with the random effect for the assessment process. While a fully parametric approach to inference would require specification of the joint distribution of the random effects $\eta_{Yi}, \eta_{Vi}$, in our semi-parametric approach we need fewer assumptions and follow Liang et al[4] in modeling this dependence through a linear link function

$$E(\eta_{Yi} \,|\, \eta_{Vi}) = \theta_0(\eta_{Vi} - 1) \tag{3}$$

for some $\theta_0$, and additionally assume that $\eta_{Vi} \sim iid\ \Gamma(1/\sigma_0^2, 1/\sigma_0^2)$. These assumptions are needed in order to specify zero mean estimating functions for $\beta$.

## 2.5 | Inference

In the case where exposure and outcome are unconfounded and we have ACAR, the causal contrast $\beta_0$ can be estimated as $E(Y_i(t)|A_i(t) = 1) - E(Y_i(t)|A_i(t) = 0)$. When we have AAR, inference can be done using doubly-weighted GEEs.[1]

Specifically, we take $\alpha_0(t) = X_t(t)\alpha_0^*$ for a deterministic function $X_t$ of $t$; Coulombe et al use a cubic spline basis. Defining $X_i^*(t) = (X_t(t), X_i(t))$ and $\beta_0^* = (\alpha_0^*, \beta_0)$, the DW-GEE solves

$$\sum_i \int_0^\tau X_i^*(t)' \frac{Y_i(t) - X_i^*(t)\beta^*}{exp(Z_{ai}(t)\hat{\gamma}_{0*})\pi(A_i(t); Z_i(t^-))} dN_i(t) = 0$$

In this paper we show how to do inference under a specific case of ANAR, that is, conditional independence of outcomes and assessment times given baseline covariates and random effects. We begin by assuming that there is no confounding. We then show how, when assessment times and exposure probabiities depend on time-varying covariates, the observation times can be thinned so that exposure and outocme are not confounded and the assessment intensity depends only on baseline covariates.

In the context of irregular observation without the additional complexity of causal inference, Liang et al[4] developed an approach to inference when assessment times and outcomes are independent given baseline covariates (which may be unmeasured). We describe it briefly, propose an extension, then discuss how multiple outputation can be used in conjunction with these models to enable causal inference.

### 2.5.1 | The Liang semi-parametric joint model

We present the Liang model as originally proposed, before considering how it could be used for causal inference. For an arbitrary row vector of covariates $X_i(t)$, a subset $W_i(t)$ of $X_i(t)$ and baseline covariates $Z_i$, we assume

$$E(Y_i(t)|X_i(t), \eta_{Yi}) = \alpha_0(t) + X_i(t)\beta_0 + W_i(t)\eta_{Yi}$$
$$\lambda_i(t) = I(C_i \geq t)\eta_{Vi}\lambda_0(t)\exp(Z_i\gamma_0),$$

where $\alpha_0$ is an unspecified function of time that is treated as a nuisance parameter. The outcome and visit processes are assumed to be conditionally independent given the random effects $\eta_{Vi}, \eta_{Yi}$ and baseline covariates $Z_i$.

Let $m_i = N_i(C_i \wedge \tau)$ denote the number of assessments for subject $i$, and note that $E(\Delta N_i(t)|m_i, \eta_{Vi}, C_i) = I(C_i \geq t)m_i\Delta\Lambda_0(t)/\Lambda_0(C_i)$, where $\Lambda_0$ is the cumulative baseline hazard. Then if we define

$$M_i(t) = \int_0^t \left( (Y_i(s) - X_i(s)\beta_0 - B_i(s)\theta_0)dN_i(s) - I(C_i \geq s)\alpha_0(s)\frac{m_i\lambda_0(s)}{\Lambda_0(C_i)}ds \right)$$

with $B_i(s) = W_i(s)E(\eta_{Vi} - 1|m_i, C_i)$, we have $E(dM_i(t)|X_i(t)) = 0$ since from Equation (3) $E(\eta_{Yi}|\eta_{Vi}) = \theta_0(\eta_{Vi} - 1)$. The Liang estimator for $(\beta_0, \theta_0)$ solves

$$\sum_{i=1}^n dM_i(t) = 0$$
$$\sum_{i=1}^n \int_0^\tau \begin{pmatrix} X_i(t) \\ \hat{B}_i(t) \end{pmatrix} dM_i(t) = 0$$

which is the same as solving

$$\sum_{i,j} \int_0^\tau \begin{pmatrix} X_i(t) - \tilde{X}(t) \\ \hat{B}_i(t) - \tilde{B}(t) \end{pmatrix} \left(Y_i(t) - X_i\beta - \hat{B}_i(t)\theta\right) dN_i(t) = 0 \quad (4)$$

where

$$\hat{B}_i(t) = \frac{(m_i - \exp(Z_i\hat{\gamma})\Lambda_0(C_i))\hat{\sigma}^2}{1 + \exp(Z_i\hat{\gamma})\hat{\Lambda}_0(C_i)\hat{\sigma}^2} W_i(t)$$
$$\tilde{X}(t) = \frac{\sum_i I(C_i \geq t)X_i(t)m_i/\hat{\Lambda}_0(C_i)}{\sum_i I(C_i \geq t)m_i/\hat{\Lambda}_0(C_i)} \qquad \tilde{B}(t) = \frac{\sum_i I(C_i \geq t)\hat{B}_i(t)m_i/\hat{\Lambda}_0(C_i)}{\sum_i I(C_i \geq t)m_i/\hat{\Lambda}_0(C_i)}$$

and $\hat{\sigma}$, $\hat{\gamma}$, $\hat{\Lambda}$ are given in the Appendix. The resulting estimates are asymptotically Normal and unbiased.[4] Although a closed-form expression for their variance exists, it is unstable in moderately sized samples and Liang et al recommend estimating standard errors via a non-parametric bootstrap.[4]

### 2.5.2 | Proposed modification to the Liang model

A disadvantage of the Liang model is that to compute $\tilde{X}_i(t)$ and $\tilde{B}_i(t)$ in (4), $X_i(t)$ must be known at every time point, regardless of whether or not a visit occurs. This is unfortunate because one of the advantages of semi-parametric models is that they can accommodate internal (endogenous) covariates, however these are often observed only at assessment times. In this section we show that if we are willing to model and estimate $\alpha_0(t)$ we need know the covariates $X(t)$ only at the assessment times. Furthermore, in some settings estimating $\alpha_0(t)$ may be of interest, for example if we wish to interpret the causal effect in terms of the mean response among the untreated (ie, a relative treatment effect); the Liang model cannot be used for this purpose because it treats $\alpha_0(t)$ as a nuisance parameter and does not estimate it.

As in the DW-GEE we take $\alpha_0(t) = X_t(t)\alpha_0^*$ for some deterministic function $X_t$ of $t$, and define $X_i^*(t) = (X_t(t), X_i(t))$ and $\beta_0^* = (\alpha_0^*, \beta_0)$. The revised model is then

$$Y_i(t) = X_i^*(t)\beta_0^* + W_i(t)\eta_{Yi} + \epsilon_i(t).$$

Taking

$$M_i^*(t) = \int_0^t (Y_i(s) - X_i^*(s)\beta_0^* - B_i(s)\theta_0)dN_i(s),$$

note that $E(dM_i^*(t)|X_i^*(t)) = 0$. Consequently we may estimate $\beta_0^*$ and $\theta_0$ by solving

$$\sum_{i=1}^{n}\int_0^\tau \begin{pmatrix} X_i(t) \\ \hat{B}_i(t) \end{pmatrix} \left(Y_i(t) - X_i^*(t)\beta^* - \hat{B}_i(t)\theta\right) dN_i(t) = 0 \tag{5}$$

These modified estimating equations do not require knowledge of $X_i(t)$ at times when there is no outcome assessment. We shall refer to the resulting estimator of $\beta^*$ as the Liang-time estimator, since it estimates the time component of the mean model for the outcomes.

Denoting $\hat{\beta}^*$ and $\hat{\theta}$ as the solution to estimating Equation (5), we show in the Appendix that $\sqrt{n}(\hat{\beta}^* - \beta_0^*)$ and $\sqrt{n}(\hat{\theta} - \theta_0)$ both converge in distribution to zero-mean Normal distributions.

### 2.5.3 | Multiple outputation to handle time-dependent covariates in the assessment intensity model

The Liang model allows us to handle dependent random effects in the assessment and outcome processes, however neither the original nor our Liang-time estimating equations can handle time-dependent covariates in the intensity model.[17] We show how multiple outputation can be used to create thinned datasets where neither the assessment intensity nor the exposure probability depends on time-dependent covariates.

Multiple outputation works by randomly discarding observations with probability inversely proportional to the observation intensity.[17] Repeating this procedure multiple times and combining the results from analysing each of the resulting thinned datasets allows for use of all the data.

Our use of outputation alters the time points at which we observe outcomes, not the population about which we make inferences. We wish to make inferences about our population had we, contrary to fact, had observation times that were independent of $Z_{ai}$ and exposures that were independent of $Z$. Since in our data the observation intensity $\lambda_i$ depends on $Z_{ai}$, certain values of these covariates are over-represented in the data. We can remedy this by, for each individual, discarding observations that have been over-sampled, so as to arrive at a dataset where $\lambda_i$ no

longer depends on the covariates. We can then further thin the data so as to remove dependence of the exposure probabilities on $Z$.

Specifically, on each outputation, we randomly sample observations with probability $\frac{s(t)}{\exp(Z_{ai}(t)\gamma_{0*})}$, where $s(t)$ is an arbitrary function of time. The resulting thinned dataset has assessment intensity

$$\lambda_i^*(t) = \lambda_i(t) \times \text{ sampling probability} = s(t)\frac{I(C_i \geq t)\eta_{Vi}\lambda_0(t)\exp(Z_{ai}(t)\gamma_{0*})}{\exp(Z_{ai}(t)\gamma_{0*})} = I(C_i \geq t)\eta_{Vi}\lambda_0^*(t)$$

for $\lambda_0^*(t) = s(t)\lambda_0(t)$. We have thus removed dependence of the assessment times on covariates $Z_{ai}$, but still have exposure probabilities that are dependent in $Z$. We thus thin the dataset again, this time sampling observations with probability $\frac{1}{\pi(A_i(t);Z_i(t^-))}$. The final, double-thinned dataset then follows the model

$$\begin{aligned} E(Y_i^a(t)|X_i(t),\eta_{Yi}) &= \alpha_0(t) + a\beta_0 + W(a,t)\eta_{Yi} \\ \lambda_i^{*a}(t) &= \lambda_i^*(t)\pi(a;Z_i(t^-)) \times \text{sampling probability} = \frac{\lambda_i^*(t)\pi(a;Z_i(t^-))}{\pi(a;Z_i(t^-))} = I(C_i \geq t)\eta_{Vi}\lambda_0^*(t) \end{aligned} \tag{6}$$

This is a special case of the Liang model and so $\beta_0$ can be estimated using the Liang estimating Equation (4) or our Liang-time estimating Equation (5).

Although we have described thinning as a two stage process, in practice it can be accomplished in a single stage. Presenting two separate steps makes it clearer how mathematically the intensity $\lambda_i^{*a}(t)$ can be made independent of both the assessment time covariates $Z_{ai}(t)$ and the exposure covariates $Z_i(t^-)$. In practice, however, the two-stage thinning presented here is equivalent to a single stage, where observations are selected with probability $\frac{s(t)}{\exp(Z_{ai}(t)\gamma_{0*})\pi(A_i(t)|Z_i(t^-))}$.

Note from Equation (6) that the observation intensity in the thinned dataset remains dependent on $\eta_{Vi}$; we have removed observations from the dataset rather than removing people. Although it is possible that in the process of any given outputation we remove all observations from a given individual, this happens independently of the random effect $\eta_{Vi}$ since we have assumed that random effects and covariates are independent. Thus the distribution of random effects in the thinned datasets is the same as in the original dataset.

In practice, $\gamma_{0*}$ and $\pi$ are not known and must be estimated, however replacing $\gamma_{0*}, \pi$ by $\hat{\gamma}_*, \hat{\pi}$ in the multiple outputation will result in estimates $\hat{\beta}$ of $\beta$ such that $\hat{\beta} - \beta$ is $o(n^{-1/2})$ provided that

$$\exp(Z_{ai}(t)\hat{\gamma}_*)\hat{\pi}(a_i(t)|Z_i(t^-)) - \exp(Z_{ai}(t)\gamma_*)\pi(a_i(t)|Z_i(t^-))$$

is $o(n^{-1/2})$.[17] If $\hat{\gamma}$ and $\hat{\pi}$ are estimated through correctly specified frailty and logistic regression models respectively, then this condition is satisfied.

To summarize, the estimation procedure proceeds as follows:

1. Estimate $\pi$ using logistic regression
2. Estimate $\gamma_*$ using a frailty model
3. Create $M$ outputted datasets where observations are selected with probability inversely proportional to $\exp(Z_{ai}(t)\hat{\gamma}_*)\hat{\pi}(a(t)|Z_i(t^-))$
4. Fit the Liang or Liang-time model to each outputted dataset to obtain estimates $\hat{\beta}^{(1)}, \hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(M)}$ of $\beta$
5. Estimate the standard errors of $\hat{\beta}^{(m)}$ for $m = 1, \ldots, M$ through a non-parametric bootstrap
6. Compute the mean of the $M$ estimates of $\beta$: $\bar{\hat{\beta}} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}^{(m)}$
7. Compute the multiple outputation variance, given by $\frac{1}{M}\sum_{i=1}^{M}\hat{var}(\hat{\beta}^{(m)}) - \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}^{(m)} - \bar{\hat{\beta}})^2$

The procedure will be asymptotically unbiased regardless of the value of $M$, however the variance of $\bar{\hat{\beta}}$ decreases with increasing $M$;[19] $M$ is thus chosen as a trade-off between computation time and estimation precision. We shall refer to multiple outputation in conjunction with the Liang model as MO-Liang and multiple outpuation in conjunction with the Liang-time model as MO-Liang-time.

## 3 | SIMULATION

We now explore the finite-sample performance of the MO-Liang and MO-Liang-time estimators, comparing these estimators to doubly-weighted GEEs in the setting where the outcome and assessment time processes are related by both observed covariates and correlated random effects.

### 3.1 | Simulation set-up

Our simulation is based on the set up of Coulombe et al,[1] with the following modifications. Since one of the benefits of a marginal structural model is to handle time-dependent confounders that are also mediators, we adapted the set-up so that the mediator of treatment effect on outcome was also a confounder. We also introduced correlated random effects in the outcome and observation models. Specifically, we took

$$\begin{aligned}
\lambda_i(t) &= 0.01\eta_{Vi}\exp(0.6A_i(t) + 0.3Z_i(t)) \\
Y_i(t) &= t + \eta_{Y1i} + (2 + \eta_{Y2i})A_i(t) - 4(Z_i(t) - E(Z_i(t)|A_i(t)) + \epsilon_i(t) \\
\eta_{Vi} &\sim Gamma(\alpha, \alpha) \\
\eta_{Y1i}|\eta_{Vi} &\sim N(\eta_{Vi} - 1, 1.8) \\
\eta_{Y2i}|\eta_{Vi} &\sim N(\eta_{Vi} - 1, 1.8) \\
\epsilon_i(t) &\sim N(\psi_i, 0.01) \\
\psi_i &\sim N(0, 0.04)
\end{aligned}$$

We took $\alpha = 100, 10, 5$, corresponding to frailty variances of 0.01, 0.1 and 0.2 respectively. The exposure $A_i(t)$ was a Bernoulli random variable with probability $\pi_i(t)$ given by

$$logit(\pi_i(t)) = 0.5 + 0.2Z_i(t-1) - 1.5A_i(t-1).$$

Finally, the mediator $Z_i(t)$ had distribution

$$\begin{aligned}
Z_i(t)|A_i(t) \sim N(2,1) \text{ if } A_i(t) = 1 \\
N(4,4) \text{ if } A_i(t) = 0
\end{aligned}$$

Conceptually, this corresponds to an exposure $A$ (for example a treatment) that is determined at the previous time point based on the previous exposure and the previous value of the mediator. The mediator is affected by the current value of the treatment, the outcome is affected by the current values of the mediator and treatment, and the observation intensity depends on the current value of the mediator and current value of the exposure. A DAG corresponding to this simulation set-up is given in Appendix Figure C1.

In the case where $\alpha = 10$, we studied two modifications, as follows. First, we introduced dependence in the mediators through random effects; we took $\eta_{Zi} \sim N(0, 0.5)$, and

$$\begin{aligned}
Z_i(t)|A_i(t), \eta_{Zi} \sim N(2 + \eta_{Zi}, 0.5) \text{ if } A_i(t) = 1 \\
N(4 + \eta_{Zi}, 3.5) \text{ if } A_i(t) = 0
\end{aligned}$$

This keeps the marginal variance of $Z_i(t)$ the same as in the original setting while inducing within-subject correlation.

Second, we allowed the observation intensity to depend on the last observed value of $Y$, specifically $\lambda_i(t) = 0.01\eta_{Vi}\exp(0.6A_i(t) + 0.3Y_i(T_{iN_i(t^-)})$.

We considered sample sizes $n = 250$ and $n = 500$. One thousand iterations were run for each data generating mechanism and for each sample size. Time-dependent random variables were simulated on a discrete grid from 0 to 2 in increments of 0.01. Outcomes $Y_i(t)$ were retained in the dataset only when $\Delta N_i(t) = 1$.

When covariates in the exposure or assessment time models are simulated from an unbounded distribution, the resulting weights must be truncated before being used for outputation. Since observations are selected with probability inversely

proportional to the observation intensity, the inverse intensity weights must be divided by a number at least as large as the largest weight. While one can use the largest estimated inverse intensity in the dataset, asymptotically this is infinite. We studied truncation at the 99th and 99.5th percentiles. Depending on the proportion of observations retained in the outputted datasets, we also studied truncation at the 95th and 99.9th percentiles.

Each simulated dataset was analysed using a DW-GEE, DW-GEE with truncation (DW-GEE-trunc), the MO-Liang estimator, and the MO-Liang-time estimator. Multiple outputations used 20 outputations. The variance of the frailty variable in the visit process model was estimated by fitting a frailty model (coxph with frailty(id)) to the full dataset rather than the method of moments approach used by Liang et al.[4] The code is available in the Appendix.

## 3.2 | Simulation results

The bias in the DW-GEE estimates increased as the variance of the frailty variable increased, with the bias in the DW-GEE-trunc estimates always exceeding that of the MO-Liang-time estimates (Table 1). Truncation increased the bias of the DW-GEE estimators, with the degree of bias increasing as the degree of truncation increased. The MO-Liang estimator had minimal bias when truncation was at the 99 $^{th}$ percentile, but became increasingly biased as the degree of truncation decreased (ie, as the truncation percentile increased). Conversely, the MO-Liang-time estimator had minimal bias when truncation was at the 99.9th percentile but became increasingly biased as the degree of truncation increased. These patterns remained consistent as the variance of the frailty variable was varied. The MO-Liang-time estimator had smaller ESEs than the MO-Liang estimator. These results held true for a sample size of 500 (see Appendix Table C1).

Similar results held when the mediators $Z$ were correlated within subjects through a random effect (see Table 2). The DW-GEE and DW-GEE-trunc estimates were all biased, regardless of truncation point, with the bias in the DW-GEE-trunc estimates exceeding that of the MO-Liang-time estimates for all truncation points. The MO-Liang estimator had small

**TABLE 1** Bias (estimated standard error) for the causal contrast in the base case, sample size of 250.

| $n = 250$ | 99th | 99.5th | 99.9th |
|---|---|---|---|
| Gamma (100,100) | | | |
| DW-GEE | 0.03 (0.44) | 0.03 (0.44) | 0.03 (0.44) |
| DW-GEE-trunc | 0.43 (0.39) | 0.28 (0.40) | 0.10 (0.43) |
| MO-Liang | 0.01 (0.51) | −0.19 (0.55) | −0.43 (0.63) |
| MO-Liang-time | 0.33 (0.45) | 0.16 (0.48) | −0.03 (0.56) |
| Gamma (10,10) | | | |
| DW-GEE | 0.11 (0.47) | 0.11 (0.47) | 0.11 (0.47) |
| DW-GEE-trunc | 0.50 (0.41) | 0.35 (0.43) | 0.18 (0.45) |
| MO-Liang | 0.00 (0.56) | −0.21 (0.60) | −0.44 (0.66) |
| MO-Liang-time | 0.31 (0.50) | 0.15 (0.53) | −0.04 (0.58) |
| Gamma (5,5) | | | |
| DW-GEE | 0.24 (0.44) | 0.24 (0.44) | 0.24 (0.44) |
| DW-GEE-trunc | 0.64 (0.39) | 0.48 (0.40) | 0.31 (0.43) |
| MO-Liang | 0.00 (0.52) | −0.20 (0.56) | −0.44 (0.64) |
| MO-Liang-time | 0.31 (0.47) | 0.15 (0.51) | −0.03 (0.57) |
| Mean no. obs | 1859 | | |
| Mean no. obs −MO (%) | 441 (24%) | 355 (19%) | 238 (13%) |

*Note*: The true value of the causal contrast is 2. Mean no. obs is the mean, over the 1000 simulated datasets, of the total number of observations in each dataset (ie, summed over individuals); Mean no. obs-MO is the mean number of observations analyzed per outputation, with the percentage of the total number of observations in parentheses.

**TABLE 2** Bias (estimated standard error) of the causal contrast when the frailty variable follows a Gamma (10,10) distribution and the mediators are dependent.

| | Truncation percentile | | |
|---|---|---|---|
| ***n* = 250** | **99th** | **99.5th** | **99.9th** |
| DW-GEE | 0.10 (0.44) | 0.10 (0.44) | 0.10 (0.44) |
| DW-GEE-trunc | 0.46 (0.39) | 0.33 (0.41) | 0.17 (0.43) |
| Mean no. obs | 1859 | | |
| MO-Liang | −0.06 (0.53) | −0.25 (0.56) | −0.46 (0.63) |
| MO-Liang-time | 0.26 (0.46) | 0.12 (0.37) | −0.05 (0.57) |
| Mean no. obs-MO (%) | 441 (24%) | 360 (19%) | 246 (13%) |

*Note*: Mean nobs is the mean number of observations per simulated dataset. Mean no. obs is the mean, over the 1000 simulated datasets, of the total number of observations in each dataset (ie, summed over individuals); Mean no. obs-MO is the mean number of observations analysed per outputation, with the percentage of the total number of observations in parentheses.

**TABLE 3** Bias (estimated standard error) of the causal contrast when the frailty follows a Gamma (10,10) distribution and the observation intensity depends on the value of the last observed outcome.

| | Truncation percentile | | |
|---|---|---|---|
| ***n* = 250** | **95%** | **99%** | **99.5%** |
| DW-GEE | 0.11 (0.92) | 0.11 (0.92) | 0.13 (0.93) |
| DW-GEE-trunc | 0.08 (0.61) | 0.09 (0.71) | 0.12 (0.76) |
| Mean no. obs | 985 | | |
| MO-Liang | −0.02 (0.87) | −0.29 (1.25) | −0.38 (1.50) |
| MO-Liang-time | −0.03 (0.85) | 0.00 (1.21) | 0.04 (1.41)* |
| Mean no. obs-MO (%) | 240 (24%) | 111 (11%) | 78 (8%) |

*81 of the 1000 datasets failed for the MO-Liang-time procedure on 99.5% truncation. Mean no. obs is the mean, over the 1000 simulated datasets, of the total number of observations in each dataset (ie, summed over individuals); Mean no. obs-MO is the mean number of observations analysed per outputation, with the percentage of the total number of observations in parentheses.

bias when weights were truncated at the 99th percentile, with this bias increasing as the degree of truncation decreased. Conversely, the MO-Liang-time estimator had small bias (−0.05) when truncation was at the 99.9th percentile with this bias increasing as the degree of truncation increased. Similar results held for a sample size of 500 (see Appendix Table C2).

When the visit process depended on the last observed outcome rather than the mediator $Z$, the bias in the DW-GEE-trunc estimator showed little change as the degree of truncation was varied, and was larger than that of the MO-Liang-time estimator at every truncation point (Table 3). The DW-GEE estimates were also more biased than the MO-Liang-time estimates. The MO-Liang-time estimator had bias less than 0.04 regardless of truncation point. Decreasing the extent of truncation continued to increase the bias of the MO-Liang estimator. Similar results held for a sample size of 500 (Appendix Table C3). Note that at a truncation percentile of 99.5% the MO-Liang-time estimator failed in 81 of the 1000 datasets. This was due to a rank deficient model matrix on solving Equation (5); specifically, the second component of $\hat{B}$ was exactly collinear with the exposure $A$ because in the outputted dataset every occurrence where $A_i(t) = 1$ corresponded to a subject for whom there was just one observation.

Across all data generating mechanisms studied, comparing across truncation points, we note that the closer the bias of the DW-GEE and DW-GEE-trunc, the smaller the bias of the MO-Liang-time estimator. The MO-Liang-time estimator acheived small bias when truncation was at the 99.9th percentile in all scenarios. Across scenarios that we examined, the MO-Liang estimator performed well provided that, on average, at least 24% of the observations were available after outputation.

# 4 | EXAMPLE: CAUSAL EFFECT OF WHEEZING ON OUTDOOR PLAY IN CHILDREN

We illustrate our methods in a study of the causal effect of wheezing on outdoor play among children aged 2–9 years enrolled in the TargetKids! study.[21] Outdoor play is associated with greater physical activity in children,[22-24] however children with respiratory conditions are advised to use caution outdoors when the air quality is poor.[25] We thus hypothesized that wheezing may reduce the amount of time that children spend playing outdoors. In this analysis our causal estimand is the mean time spent playing outdoors had all children experienced wheezing in the previous 12 months minus the mean time spent playing outdoors had no child wheezed in the previous 12 months.

TargetKids! is a prospective longitudinal cohort study that enrolls healthy children aged 0–5 years through their primary care providers and follows them as part of usual care.[21] The standard of care in Canada is for an annual well-child visit; in addition to the standard information recorded at these visits (eg, height and weight), TargetKids! caregivers complete a questionnaire capturing information on diet, exercise, and other health behaviours and exposures. In the event that the child misses their well-child visit but visits due to a health concern, the caregiver is invited to complete the questionnaire at this visit instead.

## 4.1 | Outcomes, exposures & hypothesized causal relationships

The outcome for this study is the caregiver's response to the question, "Aside from time in day care and preschool, on a typical weekday, how much time does your child spend outside in unstructured free play?". Our primary exposure is wheezing, elicited through the question, "Has your child had wheezing or whistling in the chest in the past 12 months?".[26] Secondary exposures are exercise-induced wheezing, elicited through the question, "In the past 12 months, has your child's chest sounded wheezy during or after exercise?",[26] and asthma, captured through the question "Has your child been diagnosed with asthma?" We considered data collected from January 1 2012 through May 9 2019.

As shown in Figure 2, we assume that, at any given time, changes in diet, zBMI, family income, number of siblings, maternal education, asthma diagnosis and whether or not the child lives with a smoker may be influenced by exposures at that time and not vice-versa. Temporality of the wheezing exposure supports this: wheezing captures the previous 12 months whereas other variables capture current status. It is possible that some causal relationships may be in the opposite direction to those postulated in the DAG, in particular that time spent playing outdoors might result in an episode of wheezing. This was the rationale for considering a diagnosis of asthma as a secondary exposure; previous work found that caregivers report more outdoor play in the summer months and less in the winter,[20] suggesting that relatively short windows of time are used when estimating typical outdoor play. Outdoor play is unlikely to result in a new diagnosis of asthma within a 1–2 months time frame. The rationale for considering exercise-induced wheezing as a secondary exposure is that any causal effect of wheezing on outdoor play would be expected to be larger among children who had exercise-induced wheezing.

## 4.2 | Modeling

We model the probability of each of the exposures conditionally on the outcome, exposures and other time dependent covariates at the last visit, as well as on baseline covariates. We exclude the covariates at the current visit from the exposure model as in our proposed causal framework they are mediators of the exposure-outcome relationship rather than confounders. The exposure models were fitted using a GEE, implemented using geeglm from the R package geepack.[27] We do not account for informative observation in the exposure model, as the probability weights for the exposure are conditional on being observed. Since exercise-induced wheezing was assessed only in those who reported wheezing, we fit a model for exercise-induced wheezing among those who report wheezing and calculate the probability of exercise-induced wheezing as the product of the probability of wheezing and the probability of exercise-induced wheezing given reported wheezing. Covariates were retained in the models regardless of statistical significance.

The visit process model used age as the unit of time and regressed onto baseline covariates and all the covariates listed under "Last Visit" in Figure 2, with the exception of age. As there may be clustering by region in both the outcome and the visit process,[20] the first character of the Forward Sortation Area (FSA) was included in the visit process model. Covariates

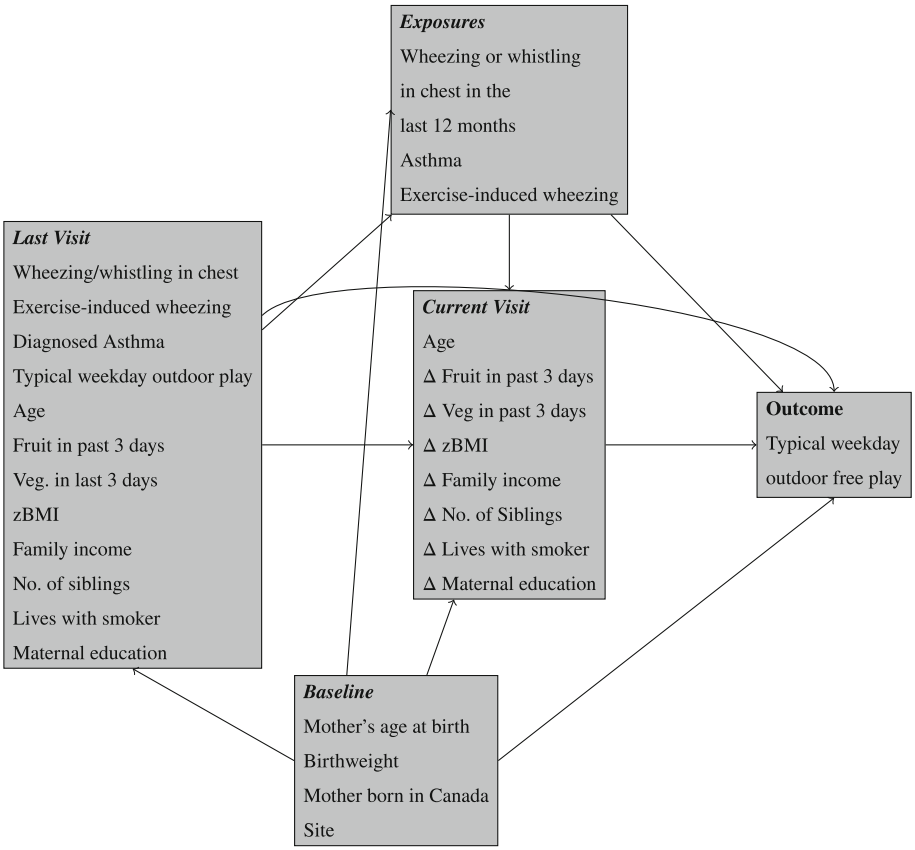


**FIGURE 2** Directed acyclic graph depicting hypothesized relationships between baseline covariates, time-varying covariates, exposures and outcome in the TargetKids! study.

were retained in the model regardless of statistical significance. The visit process model was fitted using coxph from the R package survival.[28]

We use the MO-Liang-time approach to inference. The exposures were known only at visit times, and thus the MO-Liang approach could not be used. For comparison, we also computed the unweighted GEE, doubly-weighted GEE, and MSM (ie, accounting for confounding but ignoring informative visits) estimates. Missing data was handled through multiple imputation.

## 4.3 | Results

There were 5869 children included in the sample, of whom 21% had had wheezing in the year preceding study entry, 4% reported a wheezy chest after exercise at study entry, and 7% were diagnosed with asthma at study entry (Table 4). The median follow-up time was 69 months (IQR 44 to 84) and the median number of visits over the course of follow-up was 2 (IQR 1 to 3). Figure 3 illustrates the irregularity in the visit times, and also shows that visits were more frequent around the time of the child's birthday.

The visit and exposure models are shown in Table 5. Visits were more frequent among children whose mothers were born in Canada, whose mothers held a university degree, whose mothers were 30 or over at the time the child was born as compared to in their 20 s, and who spent more time playing outdoors. Visits were less frequent among those whose family income was between $10,000 and $60,000, and who had consumed fruit in the past 3 days.

Wheezing was more common among children with very low birthweight (< 1.5 kg), who were diagnosed with asthma, whose caregiver reported wheezing and whose caregiver reported exercise-induced wheezing at the previous visit. Wheezing decreased with increasing age. Among children whose caregiver reported wheezing in the past 12 months, exercise-induced wheezing was reported more frequently in boys than girls, and in children whose mother was born in

**TABLE 4** TargetKids! cohort at baseline.

| Variable | Number (%) | Missing (%) |
|---|---|---|
| Male | 3070 (52) | 4 (0.1) |
| Diagnosed asthma | 348 (7) | 614 (11) |
| Wheezing/whistling in chest over past year | 1203 (21) | 82 (1) |
| Wheezy chest after exercise | 224 (4) | 107 (2) |
| Lives with smoker | 621 (11) | 90 (2) |
| Consumed fruit in past 3 days | 5627 (96) | 22 (0.4) |
| Consumed vegetables in past 3 days (%) | 5440 (93) | 22 (0.4) |
| Mother holds university degree | 4483 (77.5) | 86 (1) |
| Mother born in Canada | 3698 (63) | 0 |
| Age (months) (median (IQR)) | 47 (26, 64) | 0 |
| Outdoor play (minutes) (median (IQR)) | 35 (25, 60) | 388 (7) |
| Siblings (median (IQR)) | 1 (0,1) | 142 (2) |
| zBMI (median (IQR)) | 0.2 (−0.5, 0.9) | 76 (1) |
| Family income | | |
| Less than $10,000 | 72 (1) | 192 (3) |
| $10,000 to $19,999 | 114 (2) | |
| $20,000 to $29,999 | 136 (2) | |
| $30,000 to $39,999 | 166 (3) | |
| $40,000 to $49,999 | 194 (3) | |
| $50,000 to $59,999 | 197 (4) | |
| $60,000 to $79,999 | 392 (7) | |
| $80,000 to $99,999 | 519 (9) | |
| $100,000 to $149,999 | 1214 (21) | |
| $150,000 or more | 3887 (68) | |
| Mother's age at birth (years) (%) | | |
| < 20 | 30 (1) | 408 (7) |
| 20–29 | 835 (15) | |
| 30–39 | 4127 (76) | |
| 40–49 | 466 (9) | |
| ≥ 50 | 3 (0.1) | |
| Birthweight | | |
| ≤ 1.5 kg | 120 (2) | 326 (6) |
| 1.5–2.5 kg | 526 (10) | |
| > 2.5 kg | 4897 (88) | |

Canada, while it was reported less frequently in children whose mother was 40 years or older at the time of birth compared to children whose mother was in her 20 s at the time of birth. Diagnosed asthma was more common among those previously reporting a diagnosis of asthma or wheezing; we note that there were children who had a diagnosis of asthma at one visit but who did not have a diagnosis of asthma at a subsequent visit.

When weights were truncated at the 99.9th percentile, the MO-Liang-time model estimated 0.8 additional minutes of outdoor play (95% CI −3.7, 5.3) per day among children who wheezed over the preceding year compared to children who did not (Table 6). Truncation at other percentiles yielded similar results. The estimate of $\theta$ suggests that the random effects in the visit and outcome models are indeed dependent, with more frequent visits associated with less outdoor

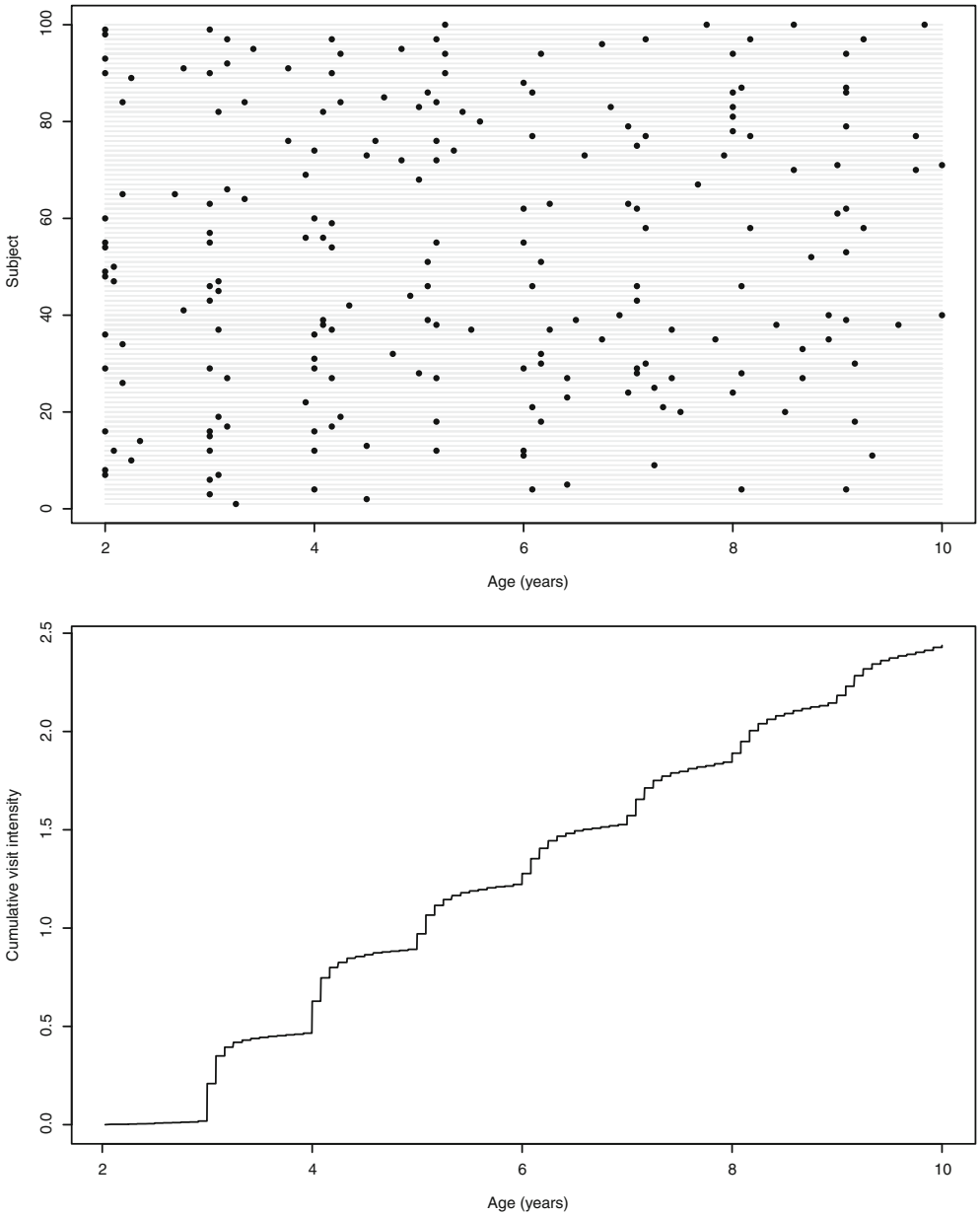


**FIGURE 3** Visit Irregularity in the TargetKids! example. The top plot is an abacus plot representing visit times for a random sample of 100 children. Each horizontal line represents a child and each point represents a visit. The bottom figure is a plot of the cumulative visit intensity as a function of age.

play: regardless of truncation point, the estimate of $\theta$ is negative and its 95% confidence interval excludes zero. When this dependence is ignored in a DW-GEE, the estimated effect of wheezing is larger; when weights were truncated at the 99.9th percentile, children who wheezed played outdoors 2.1 min more (95% CI −12, 16) than children who did not wheeze. Truncating at lower percentiles led to similar results. When confounding was accounted for but the informative visit process was ignored in an MSM, the estimated effect of wheezing was 1.1 additional minutes of outdoor play (95% CI −11, 13). When neither confounding between exposure and outcome nor the informative visit process was accounted for, children who wheezed in the preceding year played outdoors 2.0 min less (95% CI 8.9 min less to 4.9 min more) than children who did not wheeze in the preceding year.

The estimated effect of exercise-induced wheezing was an additional 2.9 min of outdoor play (95% CI 11 min less to 17 min more) when truncation was at the 99.5th percentile. When neither confounding between exposure and outcome nor the informative visit process was accounted for, children with exercise-induced wheezing played outdoors less than

**TABLE 5** Visit and exposure models for the TargetKids! dataset.

| | Intensity ratio (95% CI) | | Odds ratio (95% CI) | |
|---|---|---|---|---|
| | **Visit** | **Wheezing** | **Exercise-induced wheezing** | **Asthma** |
| At study entry | | | | |
| Male vs. Female | 1.03 (0.98, 1.09) | 1.11 (0.95, 1.29) | 1.50 (1.06, 2.12) | 1.07 (0.83, 1.38) |
| Birthweight < 1.5 kg | 1.06 (0.89, 1.26) | 1.84 (1.18, 2.89) | 0.57 (0.19, 1.69) | 1.18 (0.56, 2.48) |
| Birthweight 1.5–2.5 kg | 0.97 (0.88, 1.07) | 1.02 (0.78, 1.33) | 1.67 (0.98, 2.87) | 1.17 (0.77, 1.77) |
| Birthweight ≥ 2.5 kg | Reference | Reference | Reference | Reference |
| Mother born in Canada | 1.12 (1.06, 1.19) | 1.03 (0.86, 1.23) | 1.63 (1.07, 2.49) | 1.06 (0.77, 1.46) |
| Maternal age at birth < 20 years | 1.04 (0.67, 1.62) | 2.36 (0.63, 8.92) | 0.54 (0.03, 10.4) | 1.15 (0.14, 9.61) |
| Maternal age at birth 20–30 years | Reference | Reference | Reference | Reference |
| Maternal age at birth 30–40 years | 1.26 (1.15, 1.38) | 0.82 (0.63, 1.08) | 0.65 (0.39, 1.07) | 0.76 (0.50, 1.13) |
| Maternal age at birth ≥ 40 years | 1.21 (1.07, 1.37) | 0.87 (0.61, 1.25) | 0.40 (0.17, 0.98) | 0.67 (0.37, 1.23) |
| At last visit | | | | |
| Diagnosed asthma | 1.04 (0.93, 1.17) | 24.13 (7.7, 75.7) | 1.95 (1.31, 2.88) | 40.2 (27.7, 58.5) |
| Family income < 10K | 1.03 (0.89, 1.20) | 0.66 (0.16, 2.70) | 2.13 (0.48, 9.39) | 0.39 (0.02, 8.10) |
| Family income 10–20K | 0.39 (0.27, 0.56) | 0.64 (0.27, 1.55) | 1.97 (0.41, 9.52) | 1.17 (0.27, 5.06) |
| Family income 20–30K | 0.58 (0.45, 0.75) | 0.89 (0.41, 1.95) | 0.44 (0.06, 3.12) | 2.19 (0.92, 5.24) |
| Family income 30–40K | 0.70 (0.56, 0.88) | 0.89 (0.48, 1.64) | 0.58 (0.13, 2.57) | 1.13 (0.37, 3.49) |
| Family income 40–50K | 0.74 (0.61, 0.90) | 1.17 (0.74, 1.86) | 0.53 (0.16, 1.75) | 1.79 (0.88, 3.64) |
| Family income 50–60K | 0.74 (0.62, 0.88) | 0.89 (0.56, 1.42) | 0.12 (0.02, 0.83) | 2.06 (1.01, 4.20) |
| Family income 60–80K | 0.93 (0.79, 1.10) | 0.97 (0.68, 1.38) | 0.59 (0.24, 1.45) | 1.00 (0.48, 2.08) |
| Family income 80–100K | 0.96 (0.85, 1.08) | 0.88 (0.66, 1.18) | 0.79 (0.43, 1.45) | 0.94 (0.59, 1.49) |
| Family income 100–150K | 0.95 (0.86, 1.05) | 1.04 (0.86, 1.26) | 0.95 (0.64, 1.39) | 1.09 (0.78, 1.52) |
| Family income ≥ 150K | Reference | Reference | Reference | Reference |
| Number of siblings | 1.01 (0.95, 1.08) | 0.86 (0.77, 0.96) | 0.84 (0.66, 1.06) | 0.77 (0.64, 0.93) |
| Lives with smoker | 0.98 (0.94, 1.01) | 0.93 (0.69, 1.25) | 0.53 (0.28, 1.02) | 0.89 (0.53, 1.51) |
| Consumed fruit in past 3 days | 0.88 (0.80, 0.97) | 1.51 (0.87, 2.63) | 0.85 (0.32, 2.22) | 0.70 (0.34, 1.48) |
| Consumed vegetables in past 3 days | 1.04 (0.89, 1.21) | 1.13 (0.76, 1.67) | 0.62 (0.31, 1.25) | 1.37 (0.73, 2.55) |
| Typical weekday hours of outdoor play | 1.16 (1.02, 1.31) | 0.92 (0.84, 1.02) | 0.93 (0.78, 1.10) | 1.00 (0.86, 1.18) |
| Wheezing/Whistling in prior year | 0.97 (0.93, 1.00) | 45.7 (21.9, 95.3) | 0.58 (0.38, 0.88) | 5.50 (4.12, 7.36) |
| Exercise-induced wheezing in chest | 0.95 (0.89, 1.03) | 2.20 (1.50, 3.21) | 7.73 (5.03, 11.9) | 1.97 (1.28, 3.03) |
| zBMI | 0.98 (0.96, 1.01) | 1.08 (1.00, 1.17) | 1.09 (0.94, 1.28) | 1.06 (0.93, 1.20) |
| Mother completed university | 1.10 (1.02, 1.19) | 1.09 (0.86, 1.39) | 0.72 (0.46, 1.14) | 1.08 (0.74, 1.58) |
| Diagnosed Asthma * Wheezing | | 0.36 (0.19, 0.70) | | |
| At current visit | | | | |
| Age (years) | n/a | 0.96 (0.92, 0.99) | 1.02 (0.93, 1.11) | 1.03 (0.97, 1.10) |

**TABLE 6** Estimates of effect (95% confidence interval) of exposures (wheezing, wheezy chest on exercise, diagnosed asthma) on minutes of weekday typical outdoor free play.

| Method | Truncation | Wheezing | Exercise wheezing | Asthma |
|---|---|---|---|---|
| | | Estimate of causal effect | | |
| Unweighted | n/a | −2.0(−8.9, 4.9) | −3.3(−6.0, −0.6) | −3.2(−7.5, 1.1) |
| MSM | None | 1.1(−11, 13) | 0.5(−3.5, 4.6) | 2.6(−7, 12) |
| DW-GEE | None | 2.2(−11, 15) | 1.0(−3.0, 4.9) | 2.9(−6.3, 12) |
| DW-GEE | 99% | 2.0(−8.4, 12) | 0.7(−3.2, 4.6) | 1.3(−6.8, 9.4) |
| DW-GEE | 99.5% | 2.0(−11, 15) | 0.9(−3.0, 4.9) | 2.9(−6.0, 12) |
| DW-GEE | 99.9% | 2.1(−12, 16) | 1.0(−3.0, 4.9) | 3.2(−5.9, 12) |
| MO-Liang-time | 99% | 0.7(−3.3, 4.8) | 2.5(−8.1, 13) | 1.7(−6.7, 10) |
| MO-Liang-time | 99.5% | 1.1(−3.0, 5.3) | 2.9(−11, 17) | 3.3(−6.3, 13) |
| MO-Liang-time | 99.9% | 0.8(−3.7, 5.3) | | 3.6(−6.3, 13) |
| | | Dependence between random effects ($\theta \times 10^{-6}$) | | |
| MO-Liang-time | 99% | −1.3(−2.4, −0.3) | −1.8(−3.4, −0.2) | −2.0(−3.9, −0.2) |
| MO-Liang-time | 99.5% | −1.4(−2.5, −0.2) | −3.0(−6.0, −0.1) | −2.4(−4.8, −0.03) |
| MO-Liang-time | 99.9% | −1.4(−2.6, −0.2) | | −2.3(−5.0, 0.3) |

*Note*: DW-GEE = doubly-weighted generalized estimating equation; MO-Liang-time = multiple outputation Liang-time; MSM = marginal structural model, that is, ignoring informative visit process but accounting for confounding.

children without exercise-induced wheezing (by 3.3 min, 95% CI 0.6 less to 6 min less). Similar results held for a diagnosis of asthma.

## 5 | DISCUSSION

We have proposed two innovations to the analysis of longitudinal data with irregular assessment times. Firstly, we extended the Liang model to include estimation of the trajectory over time. Secondly, we showed how multiple outputation can be used to do causal inference when the outcome and assessment time processes include dependent random effects.

The Liang model as originally proposed treats the association between time and mean outcome as a nuisance parameter that is differenced out when constructing the estimating equations.[4] This can be helpful when the outcome of interest is the association between a covariate and outcome. In some scenarios, however, the mean trajectory over time is the target of inference (see Reference 11 for an example), and up to now the only way of handling this in the presence of dependent random effects was through fully parametric joint models. Our proposed modification to the Liang estimating equations allows for estimation of the trajectory over time while allowing for dependent random effects, but without the need to model the distribution of the outcome. There are two side-benefits to modelling the trajectory over time. Firstly, whereas the original Liang estimation procedure requires the covariates to be known at every point in time (even when there is no outcome assessment), our Liang-time estimator does not. Secondly, computation times are shorter when the trajectory over time is estimated than when it is treated as a nuisance parameter.

As in Coulombe et al[1] we have considered only a point effect of exposure, rather than a general function of the exposure history (for example, cumulative exposure over time). Often the entire exposure history will not be known in the presence of irregular observation. This is the case in the TargetKids! example, where exposures were measured only at visit times. There are however specific settings where the whole exposure history may be known; a common example is prescribed medications. Unfortunately it is not straightforward to extend our proposed multiple outputation approach in conjunction with the Liang or Liang-time models to the entire exposure history. The approach relies on defining counting processes for observed outcomes under each possible set of exposures; these are no longer well defined when the dimension of the

exposure history changes over time. Alternative approaches to handle general functions of the exposure history pose an interesting avenue for future research.

Our results show that the performance of multiple outputation in conjunction with the Liang-time model is sensitive to the truncation point. In our simulation we found that when the truncation point was selected so that the bias of the DW-GEE-trunc estimator was similar to that of the untruncated DW-GEE estimator, the bias in the MO-Liang-time estimator was small.

The poor performance of the MO-Liang estimate in the presence of lower levels of truncation was unexpected. We hypothesize that this is because the method of moments procedure to estimate $\alpha_0(s)$ becomes a poor approximation as the proportion of the data discarded on each outputation increases. It was this that motivated development of the Liang-time model, which appears to resolve the problem.

Shardell & Ferrucci[29] and Sitlani et al[30] consider joint models for outcomes and exposures linked through dependent random effects in order to handle unmeasured confounding; causal contrasts can then be identified via g-computation.[29] In the methods proposed in our paper it is possible that when the assumption of no unmeasured confounding is violated, dependence between the outcome and assessment random effects may be due to residual confounding. To investigate this, we suggest fitting a non-causal model (ie, without the exposure weights) in the first instance, and checking the estimated dependence parameter $\theta$. Dramatic changes in the estimated $\theta$ upon subsequently introducing the exposure weights may indicate the presence of unmeasured confounding.

While our proposed approach allows for dependence between the outcome and assessment time processes due to random effects, we assume that the exposures and outcomes are independent given the observed history (ie, no unmeasured confounding).

Data collected in the context of usual care is becoming increasingly available for research, creating a need to account for multiple sources of bias; co-occurrence of time-dependent confounding and irregular assessment times is one example. We have proposed methods for doing so in the presence of dependent random effects. We suggest that the presence of dependent random effects be assessed when undertaking causal inference on irregular longitudinal data in order to determine whether multiple outputation in conjunction with a semi-parametric joint model should be preferred over a doubly weighted GEE.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the TARGetKids! research network. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the TARGetKids! research network at https://www.targetkids.ca/ with the permission of the TARGetKids! research network.

## ORCID

*Eleanor M. Pullenayegum* https://orcid.org/0000-0003-4265-1330

## REFERENCES

1. Coulombe J, Moodie EEM, Platt RW. Weighted regression analysis to correct for informative monitoring times and confounders in longitudinal studies. *Biometrics*. 2021;77(1):162-174.
2. Lin H, Scharfstein D, Rosenheck R. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J R Stat Soc Ser B*. 2004;66:791-813.
3. Buzkova P, Lumley T. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Can J Stat*. 2007;35:485-500.
4. Liang Y, Lu W, Ying Z. Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*. 2009;65:377-384.
5. Sun L, Mu X, Sun Z, Tong X. Semiparametric analysis of longitudinal data with informative observation times. *Acta Math Appl Sin English Ser*. 2011;27(1):29-42.
6. Sun L, Song X, Zhou J, Liu L. Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *J Am Stat Assoc*. 2012;107(498):688-700.
7. Song X, Mu X, Sun L. Regression analysis of longitudinal data with time-dependent covariates and informative observation times. *Scand J Stat*. 2012;39(2):248-258.
8. Gasparini A, Abrams KR, Barrett JK, et al. Mixed-effects models for health care longitudinal data with an informative visiting process: a Monte Carlo simulation study. *Stat Neerl*. 2020;74(1):5-23. doi:10.1111/stan.12188
9. Ryu D, Sinha D, Mallick B, Lipsitz SL, Lipshultz S. Longitudinal studies with outcome-dependent follow-up: models and Bayesian regression. *J Am Stat Assoc*. 2007;102:952-967. doi:10.1198/00
10. Pullenayegum EM, Scharfstein DO. Randomized trials with repeatedly measured outcomes: handling irregular and potentially informative assessment times. *Epidemiol Rev*. 2022;44(1):121-137.

11. Pullenayegum EM, Lim LS. Longitudinal data subject to irregular observation: a review of methods with a focus on visit processes, assumptions, and study design. *Stat Methods Med Res*. 2016;25:2992-3014.
12. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiol*. 2000;11(5): 550-560.
13. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393-1512.
14. van der M L, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1).
15. Lin D, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *J Am Stat Assoc*. 2001;96:103-113.
16. Buzkova P, Lumley T. Semiparameteric modeling of repeated measurements under outcome-dependent follow-up. *Stat Med*. 2008;28:987-1003.
17. Pullenayegum EM. Multiple outputation for the analysis of longitudinal data subject to irregular observation. *Stat Med*. 2016;35(11):1800-1818.
18. Hoffman E, Sen P, Weinberg C. Within-cluster resampling. *Biometrika*. 2001;88:1121-1134.
19. Follmann D, Proschan M, Leifer E. Multiple outputation: inference for complex clustered data by averaging analyses from independent data. *Biometrics*. 2003;59:420-429.
20. Pullenayegum EM, Birken C, Maguire J, Collaboration TTK. Clustered longitudinal data subject to irregular observation. *Stat Methods Med Res*. 2021;30(4):1081-1100. PMID: 33509042. doi:10.1177/0962280220986193
21. Carsley S, Borkhoff C, Maguire J, et al. Cohort profile: the applied research group for kids (TARGet Kids!). *Int J Epidemiol*. 2015;44(3):776-788. doi:10.1093/ije/dyu123
22. Copeland K, Khoury J, Kalkwarf H. Child care center characteristics associated with preschoolers' physical activity. *Am J Prev Med*. 2016;50(4):470-479. doi:10.1016/j.amepre.2015.08.028
23. Tandon P, Saelens B, Zhou C, Christakis D. A comparison of preschoolers' physical activity indoors versus outdoors at child care. *Int J Environ Res Public Health*. 2018;15(11):2463. doi:10.3390/ijerph15112463
24. Razak L, Yoong S, Wiggers J, et al. Impact of scheduling multiple outdoor free-play periods in childcare on child moderate-to-vigorous physical activity: a cluster randomized trial. *Int J Behav Nutr Phys Act*. 2018;15(1):34. doi:10.1186/s12966-018-0665-5
25. AQHI Categories and Health Messages. website. 2020 http://www.airqualityontario.com/aqhi/health_messages.php#HealthMessages
26. Asher M, Keil U, Anderson H, et al. International study of asthma and allergies in childhood (ISAAC): rationale and methods. *Eur Respir J*. 1995;8(3):483-491.
27. Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. *J Stat Softw*. 2006;15(2):1-11.
28. Therneau TM. A Package for Survival Analysis in R. 2022 Website https://CRAN.R-project.org/package=survival; R package version 3.3-0
29. Shardell M, Ferrucci L. Joint mixed-effects models for causal inference with longitudinal data. *Stat Med*. 2018;37:829-846.
30. Sitlani C, Heagerty P, Blood E, Tosteson T. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Stat Med*. 2012;31:1738-1760. doi:10.1002/sim.4510
31. Lin D, Wei L, Yang L, Ying Z. Semiparametric regression for mean and rate functions of recurrent events. *J R Stat Soc Ser B*. 2000;62:711-730.


**How to cite this article:** Pullenayegum EM, Birken C, Maguire J, The TARGet Kids! Collaboration. Causal inference with longitudinal data subject to irregular assessment times. *Statistics in Medicine*. 2023;42(14):2361-2393. doi: 10.1002/sim.9727

## APPENDIX A. DETAILS FOR THE LIANG MODEL

To fit the Liang model, we require estimates of $\gamma_0$, $\Lambda_0$ and $\sigma_0$. As in Lin,[31] estimates $\hat{\gamma}$, $\hat{\Lambda}_0(t)$ are given by solving

$$\sum_i \int_0^\tau (Z_i - \tilde{Z}(t,\gamma))dN_i(t) = 0$$

$$\hat{\Lambda}_0(t;\hat{\gamma}) = \sum_i \int_0^\tau \frac{dN_i(s)}{\sum_i I((C_i > s)\exp(Z_i\gamma)} \qquad \text{where } \tilde{Z}(t;\overline{\gamma}) = \frac{\sum_i I(C_i \geq t)Z_i(t)\exp(Z_i\gamma)}{\sum_i I(C_i \geq t)\exp(Z_i\gamma)}.$$

A method of moments approach can be used to estimate $\sigma_0$.[4]

## APPENDIX B. ASYMPTOTICS FOR THE LIANG-TIME MODEL

In this section, we establish the asymptotic behavior of the estimates of $\beta^*$ on using the estimating Equation (5). The derivation is a simplified version of that in Reference 4. Letting

$$U(\beta^*, \theta) = \sum_{i=1}^{n} \int_0^{\tau} \begin{pmatrix} X_i(t) \\ \hat{B}_i(t) \end{pmatrix} \left( Y_i(t) - X_i^*(t)\beta^* - \hat{B}_i(t)\theta \right) dN_i(t)$$

we will show that $U$ can be written, to $o_p(n^{-1/2})$, as a sum of iid random variables. To do so, we use the standard asymptotic results for $\hat{\Lambda}$ and $\hat{\gamma}$, which require the following notation:

$$M_i^*(t) = N_i(t) - \int_0^t I(C_i \geq s) \exp(Z_i\gamma_0) d\Lambda_0(s)$$
$$s^{(k)}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(C_i \geq t) Z_i^k \exp(Z_i\gamma_0)$$
$$V_0 = E\left( \int_0^{\tau} \left( Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)} \right)^{\otimes 2} I(C_i \geq t) \exp(Z_i\gamma_0) d\Lambda_0(t) \right).$$

Assuming $\hat{\Lambda}$ and $\hat{\gamma}$ are estimated as in Reference 31, we then have:

$$n^{1/2}(\hat{\gamma} - \gamma_0) = V_0^{-1} n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \left( Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)} \right) dM_i^*(t) + o_{p_Z}(1)$$
$$n^{1/2}(\hat{\Lambda}(t) - \Lambda_0(t)) = n^{-1/2} \sum_{i=1}^{n} \int_0^t \frac{dM_i^*(t)(s)}{s^{(0)}(s)}$$
$$- \int_0^t \frac{s^{(1)}(s)'}{s^{(0)}(s)} d\Lambda_0(s) V_0^{-1} n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \left( Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)} \right) dM_i^*(t) + o(1)$$

where $p_Z$ is the number of covariates in $Z_i$.[31]

Furthermore,

$$n^{1/2}(\hat{\sigma}^2 - \sigma^2) = n^{-1/2} \frac{1}{T} \sum_{i=1}^{n} \left( m_i^2 - m_i - T(\sigma^2 + 1) \right) + o(1),$$

where $T = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \exp(Z_i\gamma_0)\Lambda_0(C_i)$.[4]

For an arbitrary stochastic process $F$, define

$$U(F, \beta, \theta) = \sum_{i=1}^{n} \int_0^{\tau} F_i(t)(Y_i(t) - X_i(t)\beta - \hat{B}_i(t)\theta) dN_i(t) \tag{B1}$$

and note that

$$U(F, \beta_0, \theta_0) = \sum_{i=1}^{n} \int_0^{\tau} F_i(t) dM_i(t) - \sum_{i=1}^{n} F_i(t)(\hat{B}_i(t) - B_i(t))\theta_0 dN_i(t).$$

It is straightforward to show that

$$\sqrt{n}(\hat{B}_i(t) - B_i(t)) = W_i(t) \left( \frac{(m_i - \hat{\Lambda}(C_i)\exp(Z_i\hat{\gamma}))\hat{\sigma}^2}{1 + \hat{\Lambda}(C_i)\exp(Z_i\hat{\gamma})\hat{\sigma}^2} - \frac{(m_i - \Lambda_0(C_i)\exp(Z_i\gamma_0))\sigma^2}{1 + \Lambda_0(C_i)\exp(Z_i\gamma_0)\sigma_0^2} \right)$$
$$= L_i^1(t)(\hat{\Lambda}(C_i) - \Lambda_0(C_i)) + L_i^2(t)(\hat{\sigma}^2 - \sigma_0^2) - L_i^3(t)(\hat{\gamma} - \gamma_0) + o_{p_2}(1),$$

where $p_2$ is the number of covariates in $W_i$ and

$$L_i^1(t) = W_i(t)\frac{\exp(Z_i\gamma_0)(1+m_i\sigma_0^2)\sigma_0^2}{(1+\exp(Z_i\gamma_0)\Lambda_0(C_i)\sigma_0^2)^2}$$
$$L_i^2(t) = W_i(t)\frac{m_i - \exp(Z_i\gamma_0)\Lambda_0(C_i)}{(1+\exp(Z_i\gamma_0)\Lambda_0(C_i)\sigma_0^2)^2}$$
$$L_i^3(t) = W_i(t)\frac{Z_i\exp(Z_i\gamma_0)\Lambda_0(C_i)(1+m_i\sigma_0^2)\sigma_0^2}{(1+\exp(Z_i\gamma_0)\Lambda_0(C_i)\sigma_0^2)^2}.$$

It then follows that the second term on the left-hand side of Equation (B1) can be written as

$$-\int_0^\tau H_1(F,s)\sqrt{n}d(\hat{\Lambda}(s)-\Lambda_0(s)) + H_2(F)\sqrt{n}(\hat{\sigma}^2-\sigma_0^2) - H_3(F)\sqrt{n}(\hat{\gamma}-\gamma_0) + o_f(1), \tag{B2}$$

where $f$ is the dimension of $F(t)$ and

$$H_1(F,s) = \lim_{n\to\infty}\sum_{i=1}^n\int_0^\tau I(C_i \geq s)F_i(t)\theta_0' L_i^1(t)dN_i(t)$$
$$H_j(F) = \lim_{n\to\infty}\sum_{i=1}^n\int_0^\tau F_i(t)\theta_0' L_i^j(t)dN_i(t) \text{ for } j=2,3$$

This is equal to

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\int_0^\tau H_{2i}(F,t)dM_i^*(t) - \frac{H_2(F)}{T}\left(m_i^2 - m_i - T(\sigma^2+1)\right)\right) + o_f(1),$$

where

$$H_{2i}(F,t) = \frac{H_2(F,t)}{s^{(0)}(t)} + \left(H_3(F) - \int_0^\tau H_1(F,s)\frac{s^{(1)}(s)}{s^{(0)}(s)}d\Lambda_0(s)\right)V_0^{-1}\left(Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)}\right).$$

It follows that

$$n^{-1/2}U(\beta^*,\theta) = n^{-1/2}\sum_{i=1}^n\psi_i(\beta^*,\psi,\gamma\Lambda_0,\sigma) + o_p(1),$$

where $\psi_i = (\psi_{i1},\psi_{i2})$ with

$$\psi_{i1} = \int_0^\tau X_i(t)dM_i(t) - \frac{H_2(X)}{T}(m_i^2 - m_i - T(\sigma^2+1)) + \int_0^\tau H_{2i}(X,t)dM_i^*(t)$$
$$\psi_{i2} = \int_0^\tau B_i(t)dM_i(t) - \frac{H_2(B)}{T}(m_i^2 - m_i - T(\sigma^2+1)) + \int_0^\tau H_{2i}(B,t)dM_i^*(t).$$

Since $M_i(t)$ and $M_i^*(t)$ are zero mean, and $E(m_i^2 - m_i) = T(\sigma^2+1)$, it follows that $E(\psi_i) = 0$, and so by the central limit theorem $\frac{1}{\sqrt{n}}U$ converges in distribution to a multivariate Normal with mean zero and variance $E(\psi_i\psi_i')$. Asymptotic Normality of $\hat{\beta}^*$ and $\hat{\theta}$ follows. While in principle a closed form of their variance can be derived, Liang et al note that its estimation requires plugging in estimates of the infinite-dimensional $\Lambda_0$ and is often not stable.[4] Consequently we recommend a non-parametric bootstrap for estimation of the standard errors of the Liang-time estimators.

## APPENDIX C. SUPPLEMENTARY SIMULATION INFORMATION

Figure C1 and Tables C1–C3.

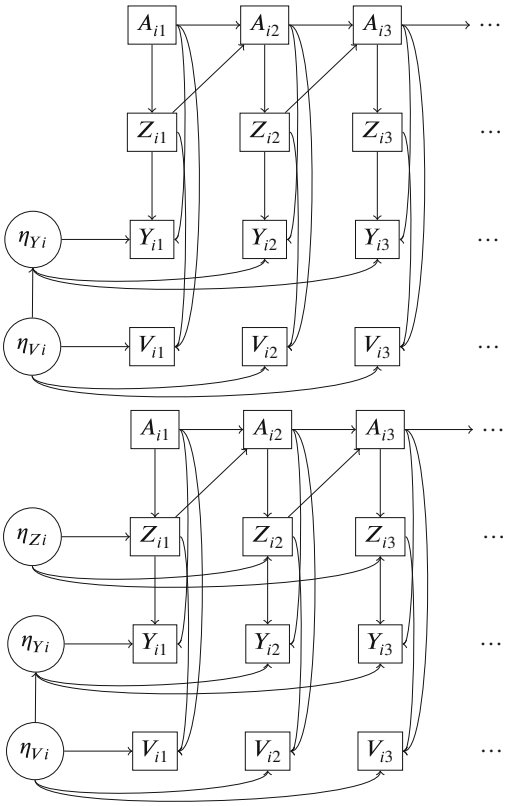


**FIGURE C1** Directed acyclic graph depicting the simulation scenarios. $Y$ is an outcome, $V$ is a visit indicator, $Z$ is a mediator and $A$ is the exposure.

**TABLE C1** Bias (estimated standard error) of the causal contrast in the base case, sample size of 500.

| $n = 500$ | 99th | 99.5th | 99.9th |
|---|---|---|---|
| Gamma (100,100) | | | |
| IIW | 0.04 (0.32) | 0.04 (0.32) | 0.04 (0.32) |
| IIW-trunc | 0.43 (0.28) | 0.28 (0.29) | 0.11 (0.31) |
| Liang | 0.03 (0.38) | −0.17 (0.41) | −0.43 (0.63) |
| Liang int | 0.33 (0.34) | 0.17 (0.37) | −0.01 (0.42) |
| Gamma (10,10) | | | |
| IIW | 0.11 (0.33) | 0.11 (0.33) | 0.11 (0.33) |
| IIW-trunc | 0.52 (0.28) | 0.37 (0.29) | 0.19 (0.31) |
| Liang | 0.00 (0.37) | −0.20 (0.41) | −0.43 (0.46) |
| Liang int | 0.32 (0.34) | 0.16 (0.36) | −0.03 (0.41) |
| Gamma (5,5) | | | |
| IIW | 0.22 (0.32) | 0.22 (0.32) | 0.22 (0.32) |
| IIW-trunc | 0.61 (0.28) | 0.46 (0.29) | 0.29 (0.31) |
| Liang | 0.00 (0.38) | −0.21 (0.41) | −0.44 (0.46) |
| Liang int | 0.30 (0.33) | 0.14 (0.36) | −0.04 (0.40) |
| Mean no. obs | 3717 | | |
| Mean no. obs −MO (%) | 877 (24%) | 702 (19%) | 459 (12%) |

*Note*: The true value is 2. Mean no. obs is the mean number of observations per simulated dataset; Mean no. obs-MO is the mean number of observations analysed per outputation, with the percentage of the total number of observations in parentheses.

**TABLE C2** Bias (estimated standard error) of the causal contrast with 500 subjects per simulated dataset when the frailty follows a Gamma(10,10) distribution and the mediators are dependent.

| $n = 500$ | 99% | 99.5% | 99.9% |
|---|---|---|---|
| IIW | 0.10 (0.32) | 0.10 (0.32) | 0.10 (0.32) |
| IIW-trunc | 0.47 (0.28) | 0.33 (0.29) | 0.17 (0.31) |
| MO-Liang | −0.05 (0.37) | −0.23 (0.40) | −0.44 (0.45) |
| MO-Liang-time | 0.27 (0.33) | 0.12 (0.25) | −0.04 (0.40) |
| Mean no. obs | 3714 | | |
| Mean no. obs-MO | 875 (24%) | 711 (19%) | 474 (13%) |

**TABLE C3** Bias (estimated standard error) of the causal contrast with 500 subjects per simulated dataset when the frailty follows a Gamma (10,10) distribution and the observation intensity depends on the last observed value of the outcome.

| $n = 500$ | 95% | 99% | 99.5% |
|---|---|---|---|
| IIW | 0.11 (0.71) | 0.11 (0.71) | 0.11 (0.71) |
| IIW-trunc | 0.07 (0.42) | 0.08 (0.50) | 0.08 (0.53) |
| MO-Liang | 0.00 (0.63) | −0.24 (0.88) | −0.30 (1.07) |
| MO-Liang-time | −0.02 (0.62) | −0.01 (0.83) | 0.02 (1.04) [+] |
| Mean no. obs | 1963 | | |
| Mean no. obs-MO | 475 (24%) | 219 (11%) | 153 (8%) |

*Note*: Mean no. obs is the mean number of observations per simulated dataset; Mean no. obs-MO is the mean number of observations analyzed per outputation, with the percentage of the total number of observations in parentheses. [+] 1 of the 1000 simulated datasets failed for the MO-Liang-time procedure on truncation at the 99.5th percentile.

## APPENDIX D. ADDITIONAL TK RESULTS

> xtable(results.irreg.unadj) > xtable(results.irreg.adj)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | | 1.4 | −0.6 | −0.6 | 1.1 | −0.56 | −0.45 | 1.1 |
| 2 | | (−9.2, 12) | (−6, 4.8) | (−6.8, 5.6) | (−5.7, 7.9) | (−6.1, 5) | (−6.8, 5.9) | (−6.5, 8.7) |
| 3 | | | | | | −235.88 | −35.82 | 3339.33 |
| 4 | | | | | | (−2513.7, 2041.9) | (−2846.6, 2774.9) | (−1020.9, 7699.5) |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | | −0.1 | −0.2 | −0.2 | −0.1 | −0.2 | −0.2 | |
| 2 | | (−0.3, 0) | (−0.3, −0.1) | (−0.3, −0.1) | (−0.3, 0) | (−0.3, −0.1) | (−0.3, 0) | (NA, NA) |
| 3 | | −13.6 | −13.7 | −13.7 | −13.8 | −14 | −14.3 | |
| 4 | | (−43.6, 16.3) | (−41.2, 13.9) | (−42.1, 14.8) | (−43.3, 15.6) | (−44.2, 16.1) | (−44.8, 16.2) | (NA, NA) |
| 5 | | −17.9 | −19 | −18.9 | −18 | −19.3 | −19.5 | |
| 6 | | (−44.6, 8.7) | (−43.8, 5.9) | (−44, 6.1) | (−43.9, 7.8) | (−47.8, 9.2) | (−48.2, 9.2) | (NA, NA) |
| 7 | | −23 | −22.2 | −22.3 | −22.9 | −22.4 | −22.7 | |
| 8 | | (−49.6, 3.6) | (−47.5, 3.1) | (−47.7, 3.1) | (−48.9, 3.2) | (−51.2, 6.3) | (−52.1, 6.7) | (NA, NA) |
| 9 | | −7.7 | −3.1 | −3.4 | −6.9 | −3.4 | −3.3 | |
| 10 | | (−20.9, 5.5) | (−14.5, 8.3) | (−14.7, 8) | (−17.8, 3.9) | (−15.9, 9.1) | (−15.9, 9.2) | (NA, NA) |
| 11 | | −2 | −0.2 | −0.1 | −2 | −0.3 | −0.1 | |
| 12 | | (−11.2, 7.2) | (−7.6, 7.3) | (−7.7, 7.6) | (−9.8, 5.9) | (−7.9, 7.3) | (−7.9, 7.7) | (NA, NA) |
| 13 | | 1.6 | 0.8 | 1.2 | 1.8 | 0.7 | 1.1 | |
| 14 | | (−11.3, 14.5) | (−5.1, 6.8) | (−5.5, 7.9) | (−5.5, 9.1) | (−5.4, 6.9) | (−5.7, 7.9) | (NA, NA) |
| 15 | | −3.3 | −3.2 | −3.1 | −3.3 | −2.8 | −3.1 | |
| 16 | | (−14.3, 7.7) | (−11.5, 5.2) | (−13, 6.7) | (−14.1, 7.5) | (−10.8, 5.3) | (−12.2, 6) | (NA, NA) |
| 17 | | 1.5 | −2.3 | −2.3 | 1 | −1.7 | −2 | |
| 18 | | (−15.6, 18.5) | (−8.2, 3.6) | (−8.3, 3.8) | (−5.8, 7.8) | (−7.8, 4.3) | (−8.1, 4.2) | (NA, NA) |
| 19 | | −3.4 | −2.5 | −2.6 | −3.4 | −2.3 | −2.4 | |
| 20 | | (−8.9, 2.2) | (−8, 2.9) | (−8, 2.9) | (−8.9, 2.1) | (−11.3, 6.8) | (−17.5, 12.8) | (NA, NA) |
| 21 | | 6.3 | 5 | 5.3 | 6.3 | 5 | 5 | |
| 22 | | (−56.9, 69.5) | (−37, 47) | (−46.3, 57) | (−56.4, 69) | (−38.5, 48.5) | (−46.9, 56.8) | (NA, NA) |
| 23 | | 1.6 | 0.3 | 0.3 | 1.5 | 0.7 | 0.6 | |
| 24 | | (−6.2, 9.4) | (−7.1, 7.8) | (−7.2, 7.9) | (−6.1, 9) | (−8.3, 9.6) | (−9.3, 10.5) | (NA, NA) |
| 25 | | −9.1 | −8 | −9 | −9 | −7.5 | −8.6 | |
| 26 | | (−34.9, 16.8) | (−24, 8) | (−26.9, 9) | (−31.2, 13.2) | (−25, 10) | (−28.3, 11) | (NA, NA) |
| 27 | | 6.1 | 3.3 | 4.1 | 6.4 | 3.6 | 4.2 | |
| 28 | | (−13.3, 25.6) | (−15, 21.7) | (−13.9, 22.1) | (−12.1, 24.9) | (−15.8, 23) | (−16.2, 24.6) | (NA, NA) |
| 29 | | 6.3 | 5 | 5.1 | 6.3 | 5.8 | 5.5 | |
| 30 | | (−32.8, 45.4) | (−31.9, 42) | (−34.2, 44.5) | (−32.9, 45.4) | (−29.9, 41.4) | (−31.2, 42.3) | (NA, NA) |
| 31 | | 8 | 8.6 | 7.7 | 7.9 | 9.3 | 8.6 | |
| 32 | | (−18.6, 34.5) | (−16.4, 33.6) | (−19.7, 35.1) | (−19.5, 35.3) | (−18.3, 36.9) | (−23.6, 40.9) | (NA, NA) |
| 33 | | 1 | −1 | −1.1 | 0.7 | −0.8 | −1.2 | |
| 34 | | (−7.9, 10) | (−6, 3.9) | (−6.5, 4.4) | (−5, 6.5) | (−5.9, 4.3) | (−6.9, 4.5) | (NA, NA) |
| 35 | | | | | | −567.4 | −591.4 | |
| 36 | | | | | | (−2755.8, 1621) | (−3232.4, 2049.6) | (NA, NA) |

## APPENDIX D. CODE

```
# input theta
# alphafn
# gamma[1], gamma[2]
# id

datagen1 <- function(id,theta,alphafn,gamma){
```

```
K1 <- rnorm(1,3,sqrt(0.5))
K2 <- rbinom(1,1,0.55)
K3 <- rnorm(1,-1.2,sqrt(0.5))

M <- rbind(c(1,0,0),c(0,1,0),c(0,0,1))
M <- diag(c(sqrt(1.8),sqrt(1.8),sqrt(1.8)))%*%M%*% diag(c(sqrt(1.8),sqrt(1.8),sqrt(1.8)))
REvec <- mvrnorm(1,rep(0,3),M)

Zfirst <- rnorm(1,4,2) + 0
pI <- 1/(1+exp(-(0.5 + 0.2*Zfirst)))
I <- rbinom(1,1,pI)
ZI1 <- rnorm(1,2,1)
ZI0 <- rnorm(1,4,2)
Z <- I*ZI1 + (1-I)*ZI0 + 0

t <- (1:200)/100
K1t <- K1 + t
K2t <- rep(K2,200)
K3errors <- rnorm(200,0,sqrt(0.005))
mat <- matrix(array(1,dim=c(200,200)),ncol=200)
mat[upper.tri(mat)] <- 0
K3t <- K3 + mat

 Zt1 <- rnorm(201,2,1)
 Zt0 <- rnorm(201,4,2)
It <- I; Zt <- Z
for(i in 1:200){
#   pIt <- 1/(1+exp(-(0.5+0.1*K1t[i]+0.05*K2t[i]-0.4*K3t[i] - 1.5*It[i])))
   pIt <- 1/(1+exp(-(0.5+0.2*Zt[i] - 1.5*It[i])))
   It <- c(It,rbinom(1,1,pIt))
 Zt <- c(Zt,It[i+1]*Zt1[i+1] + (1-It[i+1])*Zt0[i+1] + 0)

}
It <- It[-1]; Zt <- Zt[-1]

psi <- rnorm(1,0,0.2)
epsilon <- rnorm(201,0,0.1) + psi

# Dec 10th
# eta <- rgamma(1,100,100)
# Dec 10th pm
 eta <- rgamma(1,10,10)
# Dec 12th
# eta <- rgamma(1,5,5)

mu <- theta*(eta-1) + REvec[1]
muI <- theta*(eta-1) + REvec[2]

alphat <- alphafn(t)

K1t <- c(K1,K1t); K2t <- c(K2,K2t); K3t <- c(K3,K3t)
It <- c(I,It); alphat <- c(alphafn(0),alphat); Zt <- c(Z,Zt)
```

```
# Yt <- mu + alphat + (2+ muI)*It - 4*(Zt - 2*It - 4*(1-It)) + 0.4*K1 + 0.05*K2 -
0.6*K3 + epsilon
Yt <- mu + alphat + (2+ muI)*It - 4*(Zt - 2*It - 4*(1-It)) + epsilon

lambdat <- 0.01*eta*exp(gamma[1]*It + gamma[2]*Zt)
lambdat[lambdat>1] <- 1

lastZt <- c(Zfirst,Zt[-201])
datai <- cbind(rep(id,201),c(0,t),K1t,K2t,K3t,It,alphat,Zt,lastZt,Yt,lambdat)
return(datai)
}

alphafn <- function(t){return(t)}

datagen <- function(nsubj,theta,alphafn,gamma){
datalist <- lapply(1:nsubj,datagen1,theta=theta,alphafn=alphafn,gamma=gamma)
datawide <- t(matrix(unlist(datalist),ncol=201,byrow=TRUE))
data <- as.vector(datawide[,(0:(nsubj-1))*11+1])
data <- as.data.frame(data); names(data) <- "id"
data$time <- as.vector(datawide[,(0:(nsubj-1))*11+2])
data$K1 <- as.vector(datawide[,(0:(nsubj-1))*11+3])
data$K2 <- as.vector(datawide[,(0:(nsubj-1))*11+4])
data$K3 <- as.vector(datawide[,(0:(nsubj-1))*11+5])
data$I <- as.vector(datawide[,(0:(nsubj-1))*11+6])
data$alpha <- as.vector(datawide[,(0:(nsubj-1))*11+7])
data$Z <- as.vector(datawide[,(0:(nsubj-1))*11+8])
data$lastZ <- as.vector(datawide[,(0:(nsubj-1))*11+9])
data$Y <- as.vector(datawide[,(0:(nsubj-1))*11+10])
data$lambda <- as.vector(datawide[,(0:(nsubj-1))*11+11])
# names(data) <- c("id","time","K1","K2","K3","I","alpha","Z","Y","lambda")

obsgen <- function(rate){
   return(rbinom(1,1,rate))
}
data$obs <- unlist(lapply(data$lambda,obsgen))
# data$obs <- rbinom(201,1,data$lambda)
data$Y[data$obs==0] <- NA
return(data)
}

analysis.ipw.iiw <- function(data){

 data$ipw <- ipw(data)
 data$iiw <- iiw(data)
 data$weight <- data$ipw*data$iiw
 perc99 <- quantile(data$weight[data$obs==1],probs=0.99)
 data$weight.trunc <- data$weight
 data$weight.trunc[data$weight>perc99] <- perc99
 data$weight.trunc <- data$weight.trunc/mean(data$weight.trunc)
 data$mo.prob <- data$weight.trunc/max(data$weight.trunc)

 # print(summary(data$weight[data$obs==1]))
# data$weight <- data$weight/mean(data$weight,na.rm=FALSE)
```

```
 data <- data[data$obs==1,]
 tertile <- c(quantile(data$time,prob=1/3),quantile(data$time,prob=2/3))
 m <- lm(Y ~ bs(time,degree=3,knots=c(tertile)) + I,weights=weight,data=data)
 V <- vcovCL(m, cluster = data$id, type = NULL, sandwich = TRUE, fix = FALSE)
 beta <- m$coef[length(m$coef)]
 se <- sqrt(V[nrow(V),ncol(V)])
 m <- lm(Y ~ bs(time,degree=3,knots=c(tertile)) + I,weights=weight.trunc,data=data)
 V <- vcovCL(m, cluster = data$id, type = NULL, sandwich = TRUE, fix = FALSE)
 beta <- c(beta,m$coef[length(m$coef)])
 se <- c(se,sqrt(V[nrow(V),ncol(V)]))
 return(c(beta,se))
}

ipw <- function(data){
   data$lastI <- c(NA,data$I[1:(nrow(data)-1)])
   data$lastI[data$time==0] <- 0
 data$lastZ[data$time==0] <- 0
# m <- glm(I~K1 + K2 + K3 +lastI,data=data,family="binomial")
 m <- glm(I~lastZ +lastI,data=data,family="binomial")
 prob <- predict(m,type="response")
 weight <- data$I/prob + (1-data$I)/(1-prob)
 return(weight)
}

iiw <- function(data){
 m <- glm(obs ~ I + Z, data=data,family="poisson")
 intensity <- predict(m,type="response")
 m <- glm(obs ~ 1, data=data,family="poisson")
 intensity.stab <- predict(m,type="response")
 return(intensity.stab/intensity)
}

iiwmo <- function(data){
datacox <- data
datacox$tlast <- c(NA,datacox$time[1:(nrow(datacox)-1)])
datacox$tlast[datacox$time==0] <- NA
m <- coxph(Surv(tlast,time,obs)~I + Z + frailty(id),data=datacox)
datacox$lp <- drop(cbind(datacox$I,datacox$Z)
intensity <- exp(datacox$lp)

b <- basehaz(m,centered=FALSE)
Lambda0 <- max(b$haz)/2
mi <- tapply(data$obs,data$id,sum)
Lambda <- tapply(exp(datacox$lp)*0.01*Lambda0,data$id,sum)

theta <- m$history$`frailty(id)`$history[,1]
sigmasq <- theta[length(theta)]

r <- (mi + 1/sigmasq)/(Lambda+1/sigmasq)
   return(list(weight=1/intensity,sigmasq=sigmasq,eta=r))
}

analysisfn <- function(data){
```

```
   ipw.iiw <- analysis.ipw.iiw(data)
   return(ipw.iiw)
}

sim1fn <- function(it,nsubj,theta,alphafn,gamma){
 data <- datagen(nsubj=nsubj,theta=theta,alphafn=alphafn,gamma=gamma)
 res <- try(analysis.ipw.iiw(data))
 while(class(res)=="try-error"){
 data <- datagen(nsubj=nsubj,theta=theta,alphafn=alphafn,gamma=gamma)
 res <- try(analysis.ipw.iiw(data))
 }
 return(res)
}

# Proposed analysis: Liang with MO for weights

moLiang <- function(data){
 data$ipw <- ipw(data)
 i <- iiwmo(data)
 data$iiw <- i$weight
 data$weight <- data$ipw*data$iiw
 perc99 <- quantile(data$weight[data$obs==1],probs=0.99)
 data$weight.trunc <- data$weight
 data$weight.trunc[data$weight>perc99] <- perc99
 data$weight.trunc <- data$weight.trunc/mean(data$weight.trunc)
 data$mo.prob <- data$weight.trunc/max(data$weight.trunc)
 data$intercept <- 1
 data$eta <- NA
 data$eta <- i$eta[data$id]

 moest <- mo(20,Liangcausal,data=data[data$obs==1,],weights=data$weight.trunc
[data$obs==1],singleobs=FALSE,id="id",time="time",keep.first=FALSE,var=FALSE,
Yname="Y",Xnames="I",Wnames=c("intercept","I"),id.Liang="id",time.Liang="time",
Xfn=Xfn,datafull=data,sigmasq=i$sigmasq)
# data$weight <- data$weight/mean(data$weight,na.rm=FALSE)

 return(moest$est)
}

simbothfn <- function(it,nsubj,theta,alphafn,gamma){
 data <- datagen(nsubj=nsubj,theta=theta,alphafn=alphafn,gamma=gamma)
 res <- try(analysis.ipw.iiw(data))
 while(class(res)=="try-error"){
 data <- datagen(nsubj=nsubj,theta=theta,alphafn=alphafn,gamma=gamma)
 res <- try(analysis.ipw.iiw(data))
 }
 resmo <- try(moLiang(data))
 if(class(resmo)=="try-error") resmo <- rep(NA,times=5)
   resmo.mod <- try(moLiangmod(data))
 if(class(resmo.mod)=="try-error") resmo.mod <- rep(NA,times=7)
   return(list(iiwest=res,iiwmoest=resmo,iiwmoestmod=resmo.mod))
}
```

```
Xfn <- function(id,time,datafull){
   return(datafull$I[datafull$id==id & datafull$time==time])
}

Wfn <- function(id,time,datafull){
   return(c(1,datafull$I[datafull$id==id & datafull$time==time]))
}

Liangcausal <- function(data,Yname,Xnames,Wnames,id.Liang,time.Liang,Xfn,Wfn,
datafull,sigmasq){
 est <- Liangforsim(data=data,Yname=Yname,Xnames=Xnames,Wnames=Wnames,id=id.Liang,
time=time.Liang,maxfu=2,baseline=FALSE,Xfn = Xfn,Wfn = Wfn,datafull,sigmasq=sigmasq)
 return(est)
}

Liangforsim <-
function(data,datafull, Yname, Xnames, Wnames, Znames = NULL, formulaobs = NULL,
 id, time, invariant = NULL, lagvars = NULL, lagfirst = NULL,
 maxfu, baseline, n.knots = NULL, kappa = NULL, Xfn = NULL,
 Wfn = NULL,sigmasq=NULL)
{
nfull <- length(table(datafull$id)); # print("nfull"); # print(nfull)
# print("nobs post mo"); # print(sum(data$obs))
# print("nobs pre mo"); # print(sum(datafull$obs))
# print(head(data))
mjfull <- rep(0,nfull)
ids <- as.numeric(names(table(data[, names(data)
idsfull <- as.numeric(names(table(datafull[, names(datafull)
mjfull[idsfull
 id], length) - baseline
for(i in 1:nrow(datafull)){datafull$mjfull[i] <- mjfull[idsfull==datafull$id[i]]}
datafull$mX <- datafull$I*datafull$mjfull
mXbar <- tapply(datafull$mX,datafull$time,sum)
times <- as.numeric(names(table(datafull$time)))
# # print("times"); # print(times)
data$mXbar <- NA
for(i in 1:nrow(data)){data$mXbar[i] <- mXbar[times==data$time[i]]}
data$Xbar <- data$mXbar/sum(mjfull)

 if (is.null(formulaobs)) {
   fn <- function(t, tvec) return(which.min(abs(t - tvec)))
   ids <- names(table(data[, names(data)
   idnum <- array(dim = nrow(data))
   for (i in 1:nrow(data)) idnum[i] <- (1:length(ids))[data[i,
     names(data)
   if (is.data.frame(maxfu)) {
 maxfu.use <- maxfu
 for (i in 1:nrow(maxfu)) {
   maxfu.use[i, names(maxfu)
   names(maxfu)
  }
 }
 data[, names(data)
```

```
if (is.null(maxfu)) {
  maxtable <- tapply(data[, names(data)
    data[, names(data)
  maxfu.use <- cbind(1:length(maxtable), maxtable +
    max(maxtable) * 0.001)
}
n <- length(table(data[, names(data)
mi <- tapply(data[, names(data)
  id], length) - baseline
Xcols <- (1:ncol(data))[is.finite(match(names(data),
  Xnames))]
Wcols <- (1:ncol(data))[is.finite(match(names(data),
  Wnames))]
X <- array(data.matrix(data[, Xcols]), dim = c(nrow(data),
  length(Xnames)))
W <- array(data.matrix(data[, Wcols]), dim = c(nrow(data),
  length(Wnames)))
if (length(maxfu) == 1)
   maxfu.use <- cbind(idnum, rep(maxfu, length(idnum)))
maxfu.use <- maxfu.use[order(maxfu.use[, 1]), ]
data <- data[order(idnum), ]
if (length(maxfu) == 1) {
  Lambdahat <- nrow(data)/nfull
  sigmahatsq <- max((sum(mi^{2}) - sum(mi) - nfull * Lambdahat^{2})/(nfull *
    Lambdahat^{2}), 0)
    Lambdahat.scalar <- Lambdahat
  Lambdahat <- rep(Lambdahat, n)
  Ci <- rep(maxfu, n)
    # print("Lambdahat"); # print(Lambdahat)
    # print("sigma"); # print(sigmahatsq)
    if(!is.null(sigmasq)) sigmahatsq <- sigmasq
    # print("sigma"); # print(sigmahatsq)

}
if (length(maxfu) > 1) {
  maxfu.use <- maxfu.use[order(maxfu[, 1]), ]
  ids <- as.numeric(names(table(data[, names(data)
    id])))
  Ci <- as.vector(maxfu.use[order(maxfu.use[, 1]),
    2])
  data$event <- 1
  lagcols <- (1:ncol(data))[is.finite(match(names(data),
    time))]
  invarcols <- (1:ncol(data))[is.finite(match(names(data),
    id))]
  datacox <- addcensoredrows(data = data, maxfu = maxfu.use,
    tinvarcols = invarcols, id = id, time = time,
    event = "event")
  datacox <- lagfn(datacox, "time", id, time)
  formulanull <- Surv(time.lag, time, event) ~ 1
  datacox <- datacox[datacox[, names(datacox)
    time] > 0, ]
  b <- basehaz(coxph(formulanull, data = datacox))
```

```
   indexfnnocov <- function(t, time) {
     return(sum(time < t))
   }
   bindex <- sapply(Ci, indexfnnocov, time = b$time)
   bindex[bindex == 0] <- 1
   Lambdahat <- b$hazard[bindex]
   sigmahatsq <- max((sum(mi^2) - sum(mi) - nfull*(Lambdahat.scalar^2))/(nfull*
  Lambdahat.scalar^2),
     0)
     # print("sigmasq"); # print(sigmasq)
 }
 mi.Lambdahat <- mi/Lambdahat
 mi.Lambdahat[mi == 0 & Lambdahat == 0] <- 1
 W <- cbind(rep(1,nrow(data)),data$I); # print("W"); # print(head(W))
 Bhat <- array(dim = c(nrow(data), ncol(W)))
 Bbar <- Bhat
 Xbar <- array(dim = c(nrow(data), ncol(X)))
 Bmultiplier <- array(dim = nrow(data))
   Bmultiplierfull <- array(dim = nrow(datafull))
 Bmultid <- (mi - Lambdahat) * sigmahatsq/(1 + Lambdahat *
   sigmahatsq)
 Bmultidfull <- (mjfull - Lambdahat.scalar) * sigmahatsq/(1 + Lambdahat.scalar *
   sigmahatsq)
 ids <- as.numeric(names(table(data$id)))
 for (i in 1:n) Bmultiplier[data[, names(data)
  for (i in 1:nfull) Bmultiplierfull[data[, names(datafull)
 Bhat <- sweep(array(W, dim = c(nrow(data), ncol(W))),
   1, Bmultiplier, "*")
 # print("eta"); # print(summary(data$eta))
  Wfull <- cbind(rep(1,nrow(datafull)),datafull$I)
 Bhatfull <- sweep(array(Wfull, dim = c(nrow(datafull), ncol(Wfull))),
 1, Bmultiplierfull, "*")
# 1, datafull$eta-1, "*")
 mBhat <- sweep(Bhatfull,1,datafull$mjfull,"*")

 mBhatbar1 <- tapply(mBhat[,1],datafull$time,sum)
 mBhatbar2 <- tapply(mBhat[,2],datafull$time,sum)

data$mBhatbar1 <- NA
data$mBhatbar2 <- NA
times <- as.numeric(names(table(datafull$time)))
for(i in 1:nrow(data)){
 data$mBhatbar1[i] <- mBhatbar1[times==data$time[i]]
 data$mBhatbar2[i] <- mBhatbar2[times==data$time[i]]
}

data$Bbar1 <- data$mBhatbar1/sum(mjfull)
data$Bbar2 <- data$mBhatbar2/sum(mjfull)

Xbar <- data$Xbar
Bbar <- cbind(data$Bbar1,data$Bbar2)

# print("Bhat"); # print(summary(Bhat)); # print(head(Bhat))
```

```
# print("Bbar"); # print(summary(Bbar)); # print(head(Bbar))

 regX <- array((X - Xbar), dim = c(nrow(data), ncol(X)))[data[,
   names(data)
 regB <- array(Bhat - Bbar, dim = c(nrow(data), ncol(W)))[data[,
   names(data)
 # print(head(regB))
 regY <- data[, names(data)
   time] > 0]

 regpredictor <- cbind(regX, regB)
 if (sigmahatsq > 0)
   beta <- solve(t(regpredictor)
    regY)
 if (sigmahatsq == 0)
   beta <- c(solve(t(regX)
    rep(NA,ncol(W)))
 }
# print("betaLiang"); # print(beta)
   return(c(beta,sum(data$obs),sum(datafull$obs)))
}

Liangmodforsim <- function(formula,data,datafull,Wnames,id, time,baseline,
sigmahatsq)
{
 W <- cbind(rep(1,nrow(data)),data$I)

 nfull <- length(table(datafull$id))

 ids <- as.numeric(names(table(data$id)))
 n <- length(ids)

 Lambdahat <- nrow(data)/nfull
 mi <- tapply(data[, names(data)

 Bmultiplier <- array(dim = nrow(data))
 Bmultid <- (mi - Lambdahat) * sigmahatsq/(1 + Lambdahat *
   sigmahatsq)

 for (i in 1:n) Bmultiplier[data[, names(data)

 Bhat <- sweep(array(W, dim = c(nrow(data), ncol(W))),
   1, Bmultiplier, "*")

 data$Bhat1 <- Bhat[,1]
 data$Bhat2 <- Bhat[,2]

 m <- geeglm(formula=formula,data=data,id=id)
# print(summary(m))
   return(c(m$coefficients,sum(data$obs),sum(datafull$obs)))
}

Liangmodcausal <- function(data,Wnames,id.Liang,time.Liang,datafull,sigmahatsq){
```

```
 est <- Liangmodforsim(Y ~ time + I + Bhat1 + Bhat2,data=data,datafull=datafull,
 Wnames=c("Intercept","I"),id=id.Liang, time=time,baseline=FALSE,
 sigmahatsq=sigmahatsq)
 return(est)
}

moLiangmod <- function(data){
 data$ipw <- ipw(data)
 i <- iiwmo(data)
 data$iiw <- i$weight
 data$weight <- data$ipw*data$iiw
 perc99 <- quantile(data$weight[data$obs==1],probs=0.99)
 data$weight.trunc <- data$weight
 data$weight.trunc[data$weight>perc99] <- perc99
 data$weight.trunc <- data$weight.trunc/mean(data$weight.trunc)
 data$mo.prob <- data$weight.trunc/max(data$weight.trunc)
 data$intercept <- 1

 moest <- mo(20,Liangmodcausal,data=data[data$obs==1,],weights=data$weight.trunc
 [data$obs==1],singleobs=FALSE,id="id",time="time",keep.first=FALSE,var=FALSE,
 Wnames=c("intercept","I"),id.Liang="id",time.Liang="time",datafull=data,
 sigmahatsq=i$sigmasq)

 return(moest$est)
}
```

---

```
# input theta
# alphafn
# gamma[1], gamma[2]
# id

library(splines)
library(geepack)
library(lme4)
library(IrregLong)
library(sandwich)
library(coxme)
library(MASS)

gamma <- c(0.6,0.3)
theta <- 1
n <- 250

source("~/ObsPatientVisit/Causal/causalmocore.depZ.indRE.R")

set.seed(3010311)
res <- lapply(1:100,simbothfn,theta=theta,alphafn=alphafn,gamma=gamma,nsubj=250)
dput(res,"res1.Robject")
```

---