Understanding and addressing the bias due to covariate-driven observation times in longitudinal observational studies

Janie Coulombe
janie.coulombe@mail.mcgill.ca

Joint work with Dr Erica E.M. Moodie and Dr Robert W. Platt
McGill University

Seminar presented in the Department of Statistics
NC State University
October 19, 2021

Background

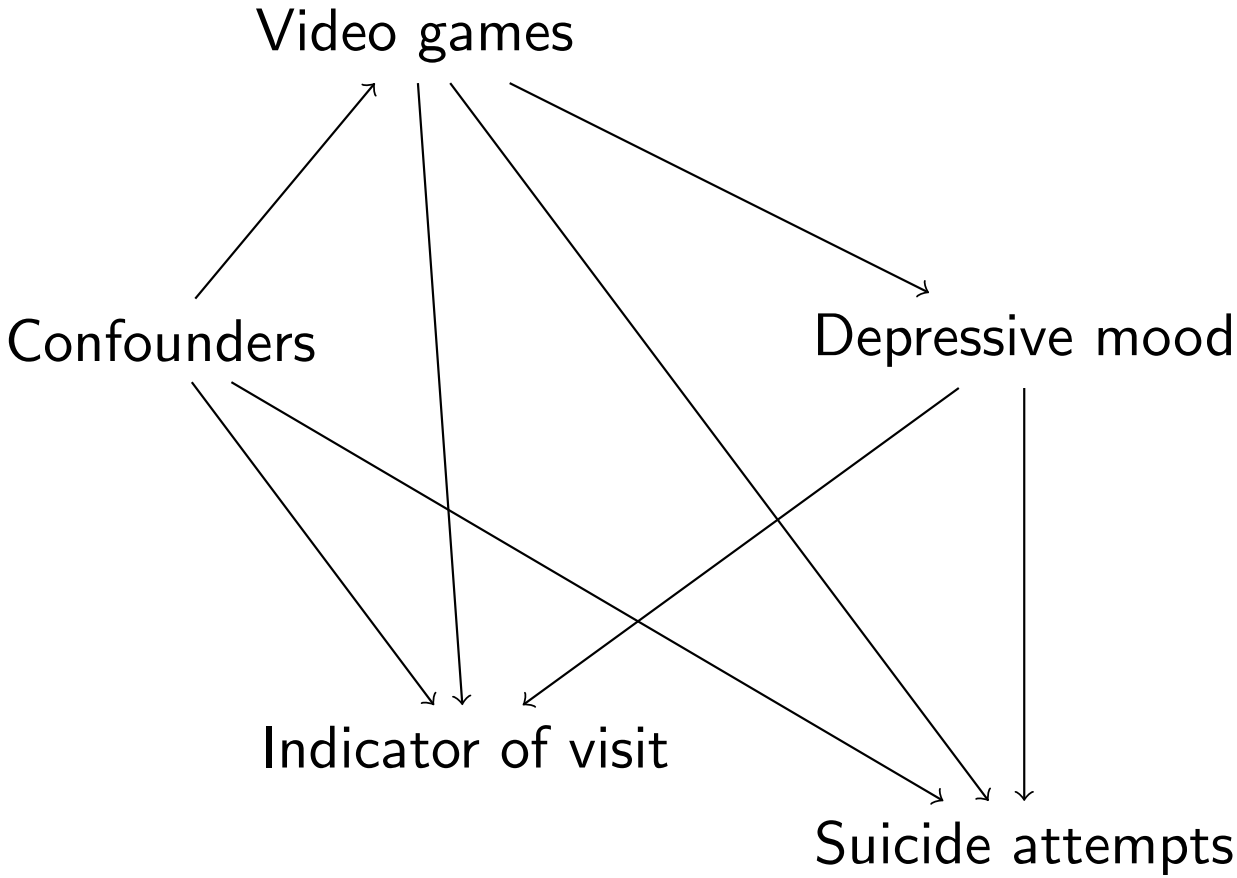
- Abundance of data not meant for research purposes
- Electronic health records (EHR) data contain rich, longitudinal information (repeated clinical measurements)
- Useful for answering causal inference questions when an experiment is not feasible
- Focus on the estimation of the marginal effect of treatment (or exposure)

Background cont'd

- However, the effects can be "distorted":
  - Confounding
  - Covariate-driven monitoring times (CDMT) of a clinical measurement
  - Missing data
  - Measurement error
- Bias due to CDMT: in our case due to collider stratification bias (Greenland, 2003)
- It could be due to, e.g., confounding by the observation schedule, unmeasured variable linking the observation and the outcome processes, etc.

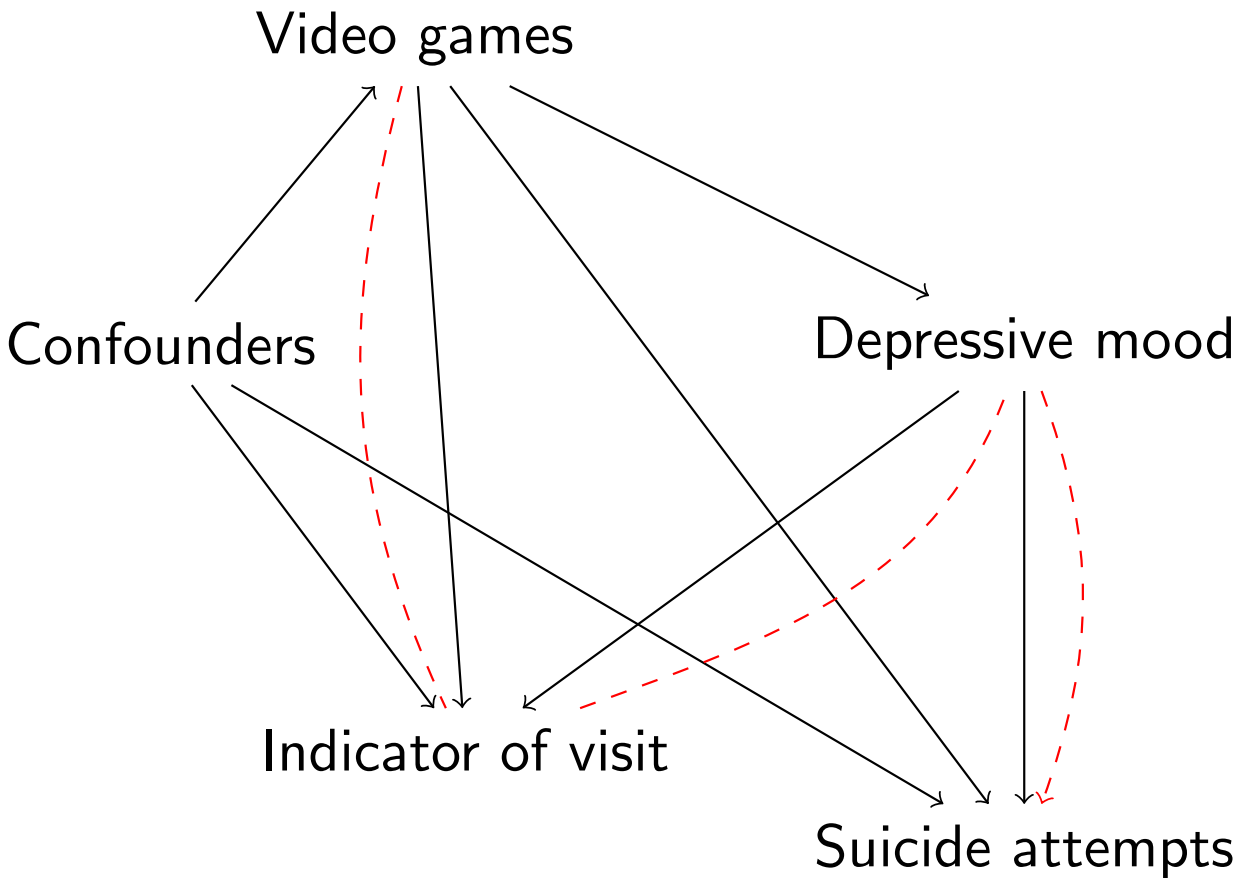
Remark. Drawing the assumed data generating mechanism (causal diagram) helps in finding the biasing paths to block

Example 1: Time spent playing video games, Coulombe et al. (2021)



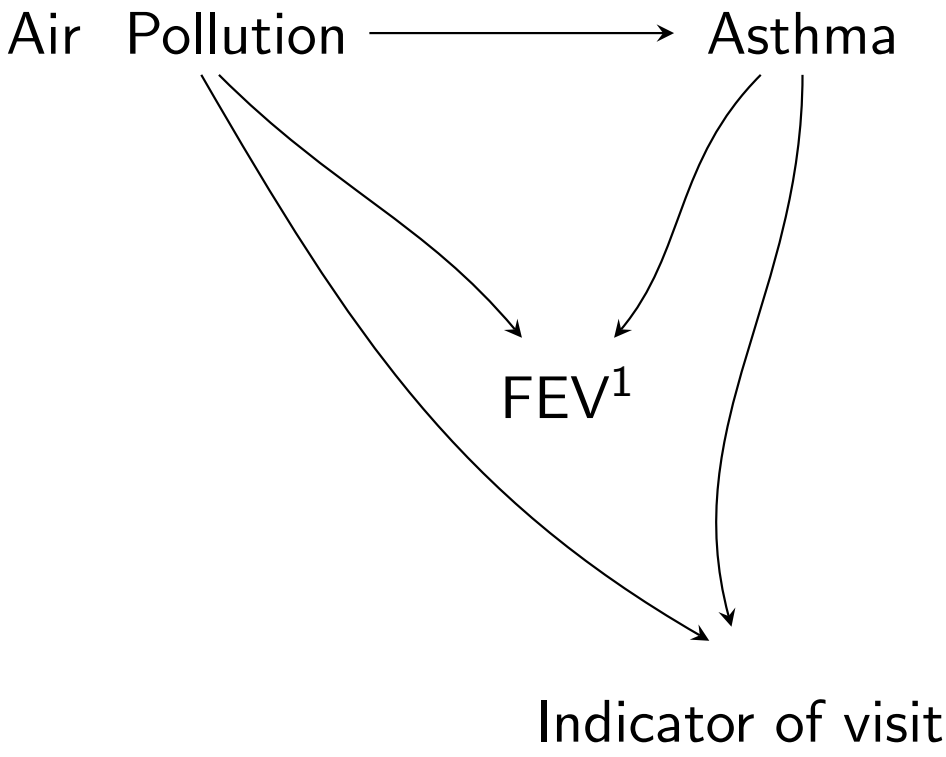
Note. Observation times, monitoring times, visit times $\cong$ times when the outcome is observed.

Example 1: Time spent playing video games, Coulombe et al. (2021)



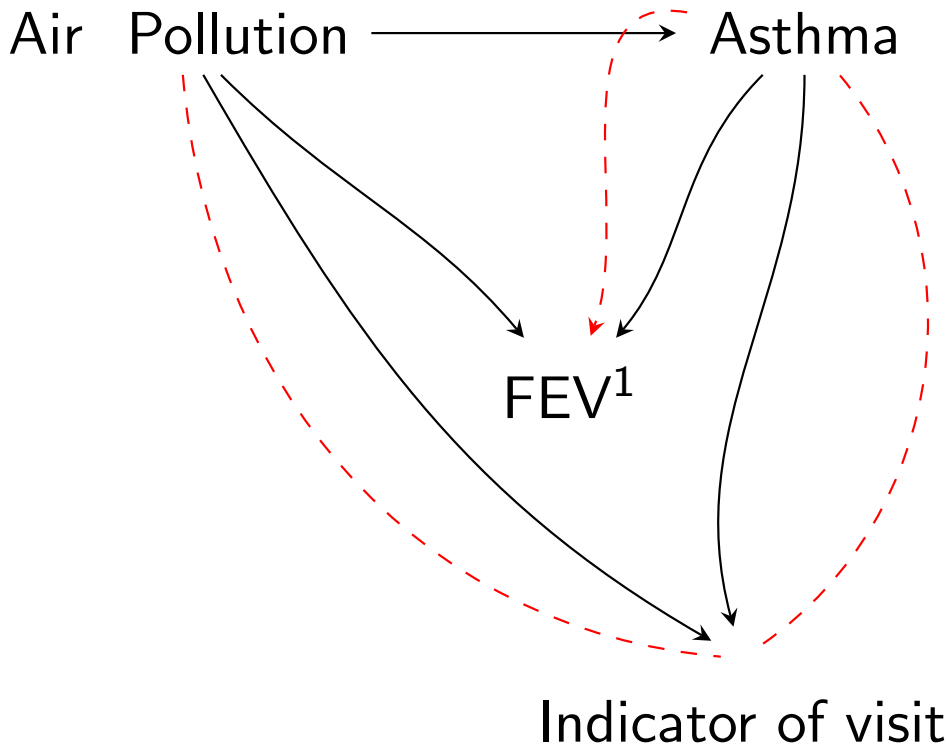
Note. Observation times, monitoring times, visit times $\cong$ times when the outcome is observed.

Example 2: Air pollution, Buzkova et Lumley (2007)



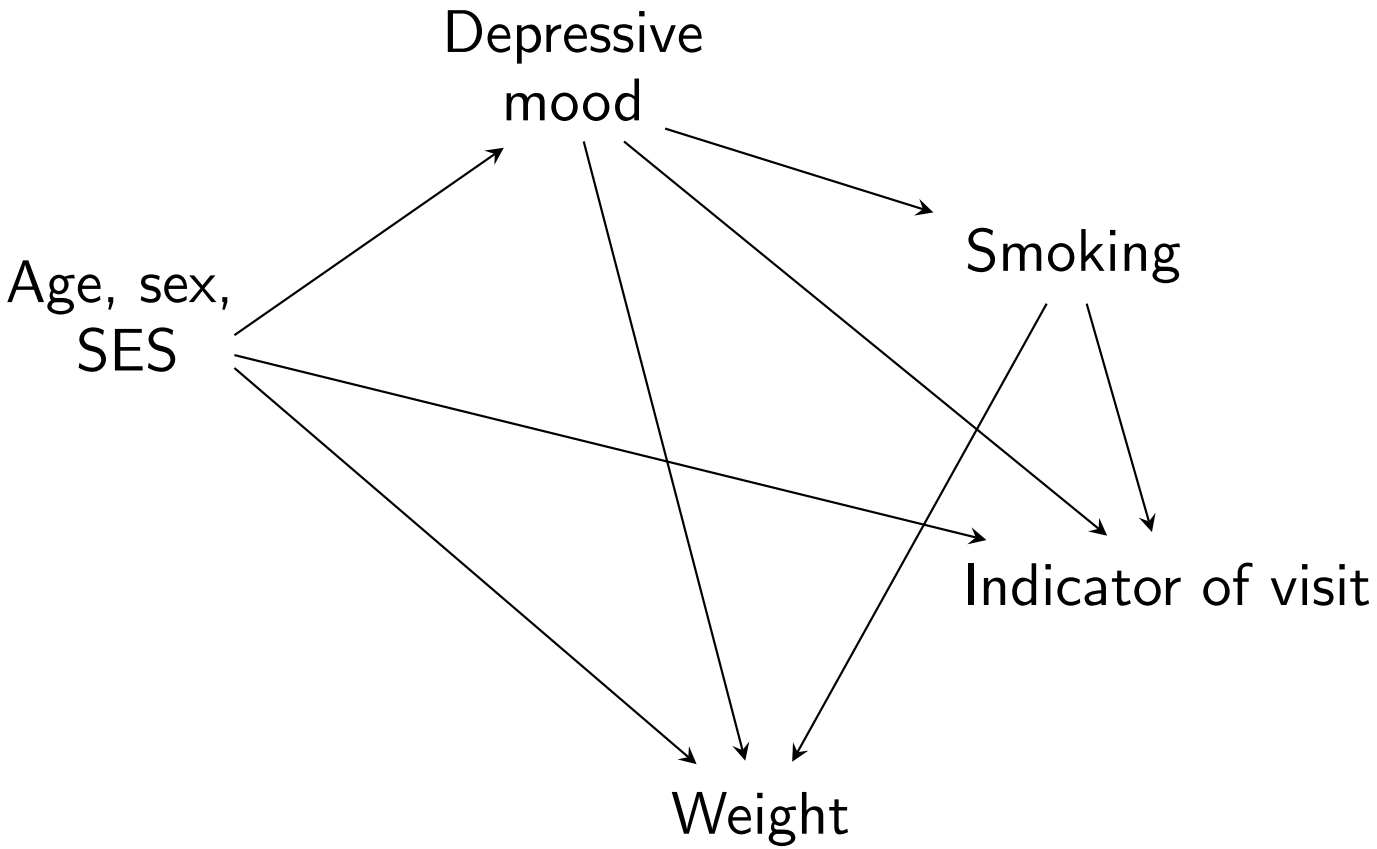
1. Forced expiratory volume

Example 2: Air pollution, Buzkova et Lumley (2007)

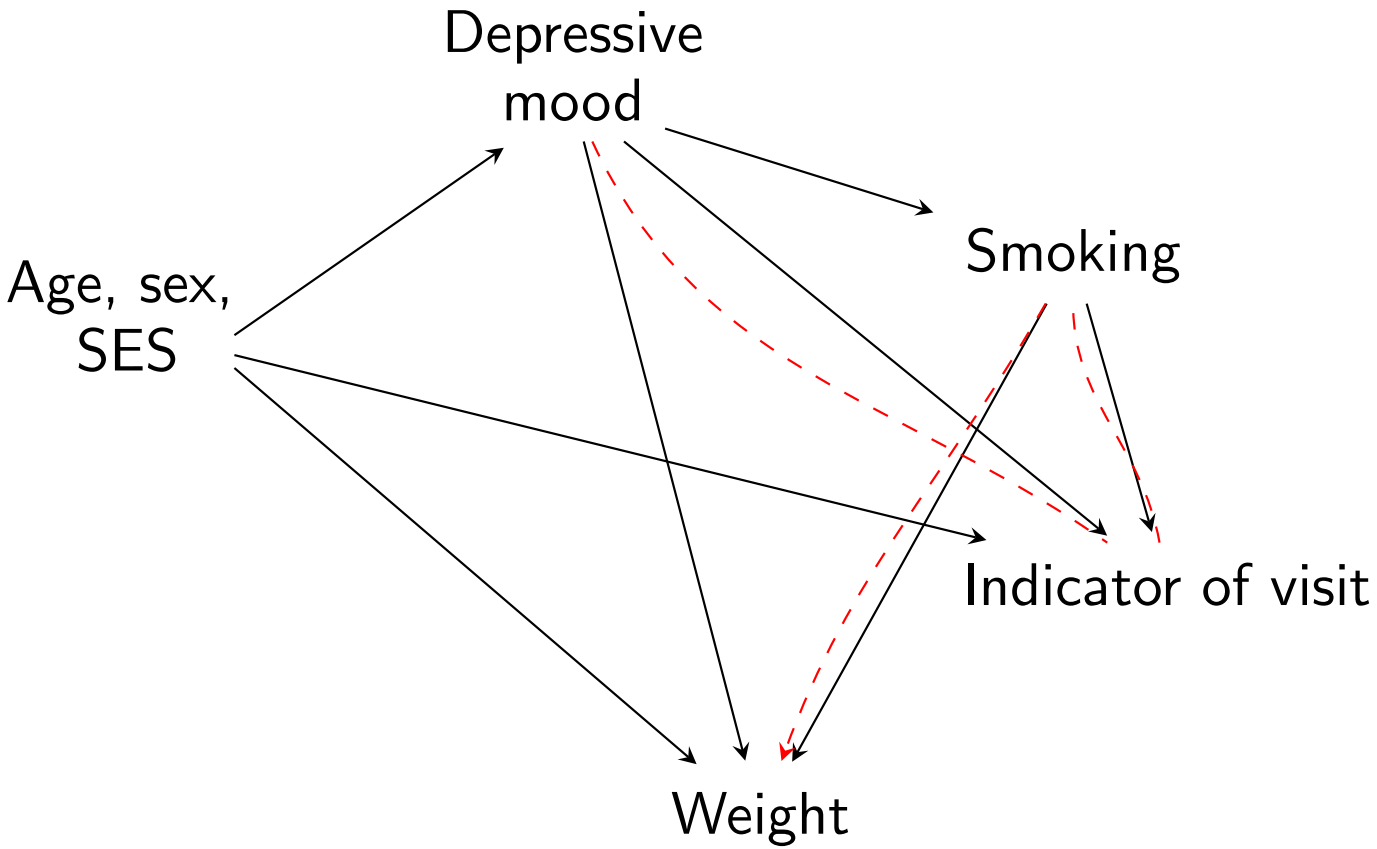


1. Forced expiratory volume

Example 3: Depressive mood, Coulombe et al. (2021)



Example 3: Depressive mood, Coulombe et al. (2021)



Methods proposed for irregular observation times

In the statistical literature, broadly categorized as

- Fully parametric specification (see, e.g., Lipsitz et al., 2002)
- Joint models (JM)
- Weighted least squares (WLS) estimators with inverse intensity of visit (IIV) weights (Lin et al., 2004)

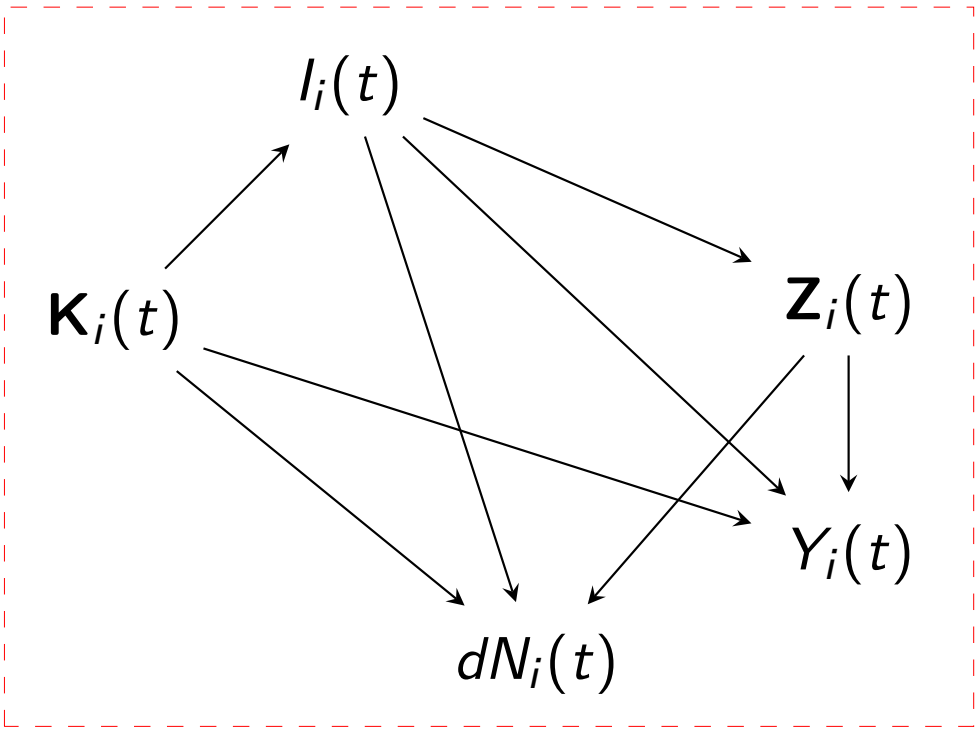
A great review is found in Pullenayegum and Lim (2016)

Question: In a causal inference framework, can we use JM/latent variables?

Remark: WLS estimators are appealing in such a context

New estimators

Notation, individual i at time t



$I_i(t)$ the exposure

$Y_i(t)$ the outcome

$\mathbf{K}_i(t)$ the confounders

$\mathbf{Z}_i(t)$ the mediators

$dN_i(t)$ a monitoring indicator

$$\mathbf{V}_i(t) = \{\mathbf{Z}_i(t), I_i(t), \mathbf{K}_i(t)\}$$

$0 \text{ — } 1 \text{ — } 2 \quad \cdots \quad t \quad \cdots \quad C_i \leq \tau$

$$\xi_i(t) = \mathbb{I}(C_i \geq t)$$

Inference

We use the potential outcome framework (Neyman, 1923; Rubin, 1974) to express the estimand of interest, given by

$$E[Y_{i1}(t) - Y_{i0}(t)]$$

for $Y_{i1}(t)$ the outcome that would have been observed at time t, in individual i, was he given treatment 1 and $Y_{i0}(t)$, the treatment 0.

Causal assumptions: Conditional exchangeability, positivity of treatment and monitoring, and consistency.

Inference cont'd

Lin and Ying (2001) and Buzkova and Lumley (2009) used the following marginal model to develop an estimator for the marginal effect of variables in their design matrix $\mathbf{X}_i(t)$:

$$E[Y_i(t)|\mathbf{X}_i(t)] = \alpha(t) + \boldsymbol{\beta}'\mathbf{X}_i(t),$$

which in our case corresponds to

$$E[Y_i(t)|I_i(t)] = \alpha(t) + \beta_I I_i(t)). \quad \text{(O1)}$$

There, $\alpha(t)$ is an arbitrary function of time, $\mathbf{Y}(t)$ a continuous longitudinal outcome and $\mathbf{I}(t)$ the exposure at time t. The model (O1) can be used to estimate

$$E[Y_{i1}(t) - Y_{i0}(t)].$$

Inference cont'd

However, an estimator for β in (O1) is not appropriate in a context where there are imbalances between treatment groups, due to confounders.

For now, we thus focus on the conditional model

$$E[Y_i(t)|I_i(t), \mathbf{K}_i(t)] = \alpha(t) + \beta_I I_i(t) + \boldsymbol{\beta}_{\mathbf{K}} \mathbf{K}_i(t). \qquad \text{(O2)}$$

Later we will see how to get an estimate for the marginal effect of treatment in a *pseudo-population* with no imbalance due to visit process and confounders.

Lin and Ying's estimator

Lin and Ying proposed an extension of ordinary least squares that considers CDMT, based on the zero-mean stochastic process

$$M_i(t; A, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \int_0^t \left\{ Y_i(s) - \boldsymbol{\beta}' \mathbf{X}_i(s) \right\} dN_i(s) - \xi_i(s) \exp\left\{ \boldsymbol{\gamma}' \mathbf{Z}_i(s) \right\} dA(s),$$

$i = 1, ..., n$, with $A(t) = \int_0^t \alpha(s) d\Lambda_0(s)$.

Assumptions

- $\mathbf{X}_i(t)$ and $\mathbf{Z}_i(t)$ are collected at all times $0 \leq t \leq \tau$
- $\mathbf{Z} \subset \mathbf{X}$ the subset that affects visit times
- Censoring is not informative
- Proportional rate model $\mathbb{E}\left[dN_i(t) | \mathbf{Z}_i(t)\right] = \exp\left\{\boldsymbol{\gamma}' \mathbf{Z}_i(t)\right\} d\Lambda_0(t)$

Lin and Ying's estimator cont'd

$$\hat{\boldsymbol{\beta}}_{Lin} = \left[\sum_{i=1}^{n} \int_0^{\infty} W(t) \overbrace{\left\{ \mathbf{X}_i(t) - \bar{\mathbf{X}}(t;\hat{\gamma}) \right\}^{\otimes 2}}^{\text{“}\mathbf{X}^T\mathbf{X}\text{”}} dN_i(t) \right]^{-1}$$

$$\times \sum_{i=1}^{n} \int_0^{\infty} W(t) \overbrace{\left\{ \mathbf{X}_i(t) - \bar{\mathbf{X}}(t;\hat{\gamma}) \right\} \left\{ Y_i(t) - \bar{Y}^*(t;\hat{\gamma}) \right\}}^{\text{“}\mathbf{X}^T\mathbf{Y}\text{”}} dN_i(t)$$

with

$$\bar{\mathbf{X}}_p(t;\gamma) = \frac{\sum_{i=1}^n \xi_i(t) \exp^{\gamma' \mathbf{Z}_i(t)} \mathbf{X}_{ip}(t)}{\sum_{j=1}^n \xi_j(t) \exp^{\gamma' \mathbf{Z}_j(t)}}, \bar{Y}^*(t;\gamma) = \frac{\sum_{i=1}^n \xi_i(t) \exp^{\gamma' \mathbf{Z}_i(t)} Y_i^*(t)}{\sum_{j=1}^n \xi_j(t) \exp^{\gamma' \mathbf{Z}_j(t)}},$$

$Y^*(t)$ a nearest-neighbor approximation to $Y(t)$ and $W(t)$ a time-dependent weight; here we use $W(t) = 1$.

Buzkova and Lumley's estimator

The previous authors make the strong assumption that $\mathbf{Z}_i(t)$ are included in $\mathbf{X}_i(t)$, excluding the possibility for $\mathbf{Z}_i(t)$ to be, e.g., mediators.

Buzkova and Lumley extend this by including a "rate ratio" weight to Lin and Ying's estimator, given by

$$\rho_i(t;\gamma,\delta) = \frac{\exp\{\gamma'\mathbf{V}_i(t)\}}{\exp\{\delta'\mathbf{X}_i(t)\}}$$

with $\mathbf{V}_i(t)$ containing all variables that affect visit times (e.g., $\mathbf{V}_i(t) = \{\mathbf{Z}_i(t), \mathbf{X}_i(t)\}$).

Buzkova and Lumley's estimator cont'd

We now assume $E[dN_i(t)|\mathbf{V}_i(t)] = \xi_i(t)\exp(\gamma'\mathbf{V}_i(t))\,d\Lambda_0(t)$ and our inverse rate ratio is given by

$$\rho_i(t;\gamma) = \frac{\exp(\gamma_1'\mathbf{K}_i(t) + \gamma_2'\mathbf{Z}_i(t) + \gamma_3 I_i(t))}{\exp(\gamma_I I_i(t))}.$$

Using a similar method as Buzkova and Lumley we obtain the following estimator (consistent for the marginal effect when there is no confounding):

$$\hat{\beta}_{BL} = \left[\sum_{i=1}^{n}\int_0^\tau \frac{W(t)}{\rho_i(t;\gamma)}\left(I_i(t) - \bar{I}(t;\gamma_I)\right)^2 dN_i(t)\right]^{-1}$$
$$\times \sum_{i=1}^{n}\int_0^\tau \frac{W(t)}{\rho_i(t;\gamma)}\left(I_i(t) - \bar{I}(t;\gamma_I)\right)\left(Y_i(t) - \bar{Y}^*(t;\gamma_I)\right) dN_i(t).$$

IPCTM estimator

We propose to incorporate an adjustment for confounders.

We try incorporating a standard inverse probability of treatment weight, which leads to more bias than we expected.

If we focus on the conditional effect of treatment (model O2) we obtain the following estimators:

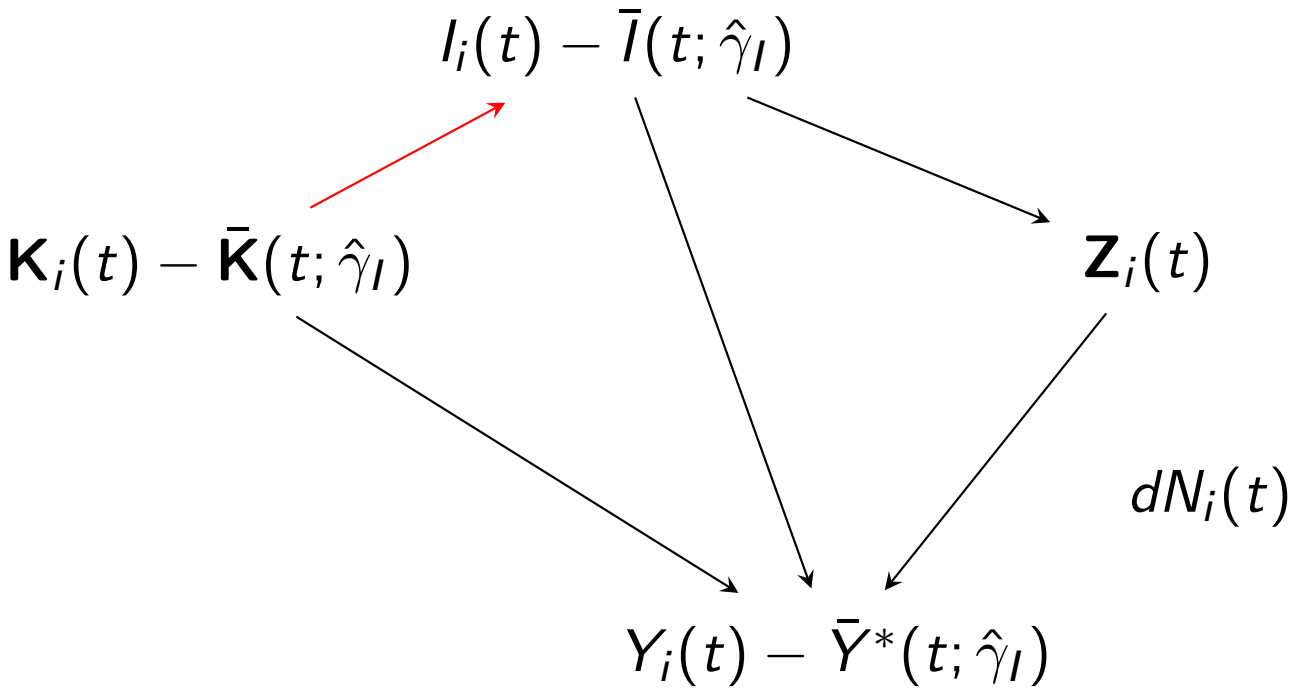
$$[\hat{\beta}_I \ \hat{\beta}_{\mathbf{k}}]' = \left[\sum_{i=1}^{n} \int_0^{\tau} \frac{W(t)}{\rho_i(t;\hat{\gamma})} \begin{pmatrix} I_i(t) - \bar{I}(t;\hat{\gamma}_I) \\ \mathbf{K}_i(t) - \bar{\mathbf{K}}(t;\hat{\gamma}_I) \end{pmatrix}^{\otimes 2} dN_i(t)\right]^{-1}$$
$$\times \sum_{i=1}^{n} \int_0^{\tau}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

IPCTM estimator cont'd

Note, estimating the marginal effect we are interested in is equivalent to estimating β in

$$E[Y_i(t) - \bar{Y}^*(t;\hat{\gamma}_I)|I_i(t) - \bar{I}(t;\hat{\gamma}_I)] = \beta\left\{I_i(t) - \bar{I}(t;\hat{\gamma}_I)\right\}.$$

The causal DAG is now equivalent to



IPCTM estimator cont'd

Generalization of the IPT weight:

$$E\left[I_i(t) - \bar{I}(t;\hat{\gamma}_I)|\mathbf{K}_i(t) - \bar{\mathbf{K}}(t;\hat{\gamma}_I)\right] = \psi_0 + \psi_{\mathbf{1}}{}'(\mathbf{K}_i(t) - \bar{\mathbf{K}}(t;\hat{\gamma}_I)), \quad (1)$$

$$E\left[I_i(t) - \bar{I}(t;\hat{\gamma}_I)\right] = \psi_m. \quad (2)$$

The stabilized generalized weight is provided by

$$\text{sgw}_i(t;\hat{\psi}) = \frac{g^{-1}\left(\hat{\psi}_0 + \hat{\psi}_{\mathbf{1}}'(\mathbf{K}_i(t) - \bar{\mathbf{K}}(t;\hat{\gamma}_I))\right)}{g^{-1}\left(\hat{\psi}_m\right)}$$

where

$$g^{-1}(\hat{a}_i(t)) = 1/\sqrt{2\pi\hat{\sigma}_a^2}\exp\left(-\hat{\epsilon}_{a,i}(t)^2/(2\hat{\sigma}_a^2)\right)$$

is the Normal density evaluated at $\hat{\epsilon}_{a,i}(t) = \left(I_i(t) - \bar{I}(t;\hat{\gamma}_I) - \hat{a}_i(t)\right)$ and $\hat{\sigma}_a^2 = var(\hat{\epsilon}_{a,i}(t))$ (Robins et al., 2000).

IPCTM estimator cont'd

The *inverse probability of centered treatment and monitoring* (*IPCTM*) weighted estimator is given by

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

This estimator accounts for informative monitoring times and confounding. It is also semiparametric and does not require the estimation of $\alpha(t)$ in (O1).

FIPTM estimator

Second proposed estimator: the flexible inverse probability of treatment and monitoring weighted estimator ($\hat{\beta}_{FIPTM}$), which rather models $\alpha(t)$ using cubic splines. Compute the vector of coefficients

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

with $\mathbf{S}_i(t)$ a design matrix that contains the intervention and the cubic spline terms, $\mathrm{e}_i(t;\boldsymbol{\omega})$ a standard IPT weight, and

$$\frac{1}{\varphi_i(t;\boldsymbol{\gamma})} = \frac{1}{\exp\left(\boldsymbol{\gamma}_1'\mathbf{K}_i(t) + \boldsymbol{\gamma}_2'\mathbf{Z}_i(t) + \gamma_3 l_i(t)\right)}.$$

The treatment coefficient is further denoted by $\hat{\beta}_{FIPTM}$.

Simulation setting

Time-fixed treatment:

$$K_i = \{K_{1i}, K_{2i}, K_{3i}\} \sim (N(1,1), Bern(0.55), N(0,1))$$

$$I_i \sim Bern(p_i) \text{ with } p_i = \frac{\exp(0.5 + 0.8K_{1i} + 0.05K_{2i} - 1K_{3i})}{1 + \exp(0.5 + 0.8K_{1i} + 0.05K_{2i} - 1K_{3i})}$$

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

$$Y_i(t) = \alpha_0(t) + 1\, I_i + 3\,\{Z_i(t) - E\,[Z_i(t)|I_i]\} +$$
$$0.4K_{i1i} + 0.05K_{2i} - 0.6K_{3i} + \epsilon_i(t)$$
$$\text{with } \epsilon_i(t)|\phi_i \sim N(\phi_i, 0.01), \phi_i \sim N(0, 0.04)$$

Simulation setting cont'd

The observation times are simulated according to a visit rate at time t defined as

$$\lambda_i(t) = \eta_i \exp\{\gamma_I I_i(t) + \gamma_z Z_i(t)\}$$

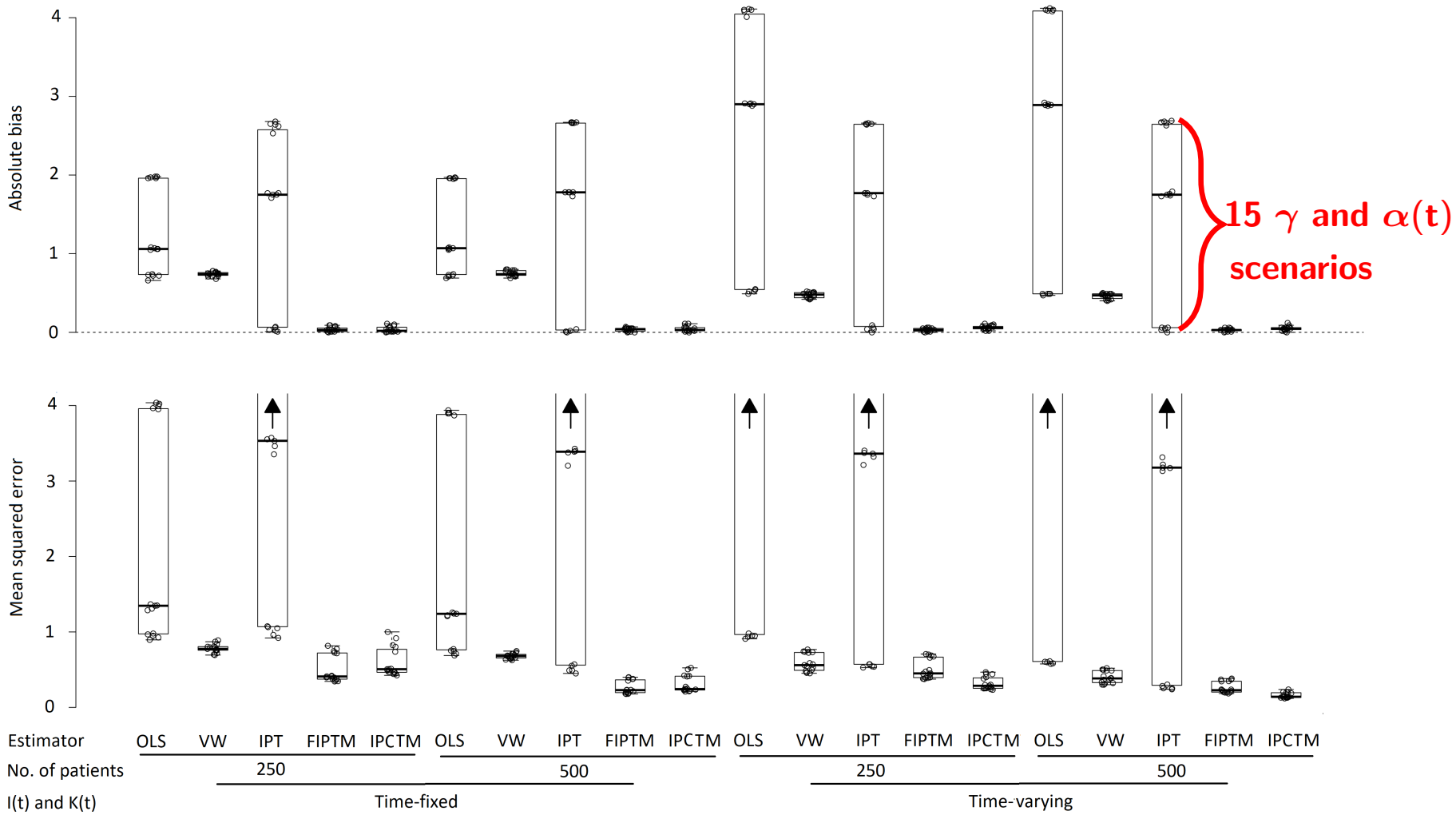
with η_i a random effect for each individual and $\boldsymbol{\gamma} = (\gamma_I, \gamma_Z)$ different "strengths".

Time-varying treatment:

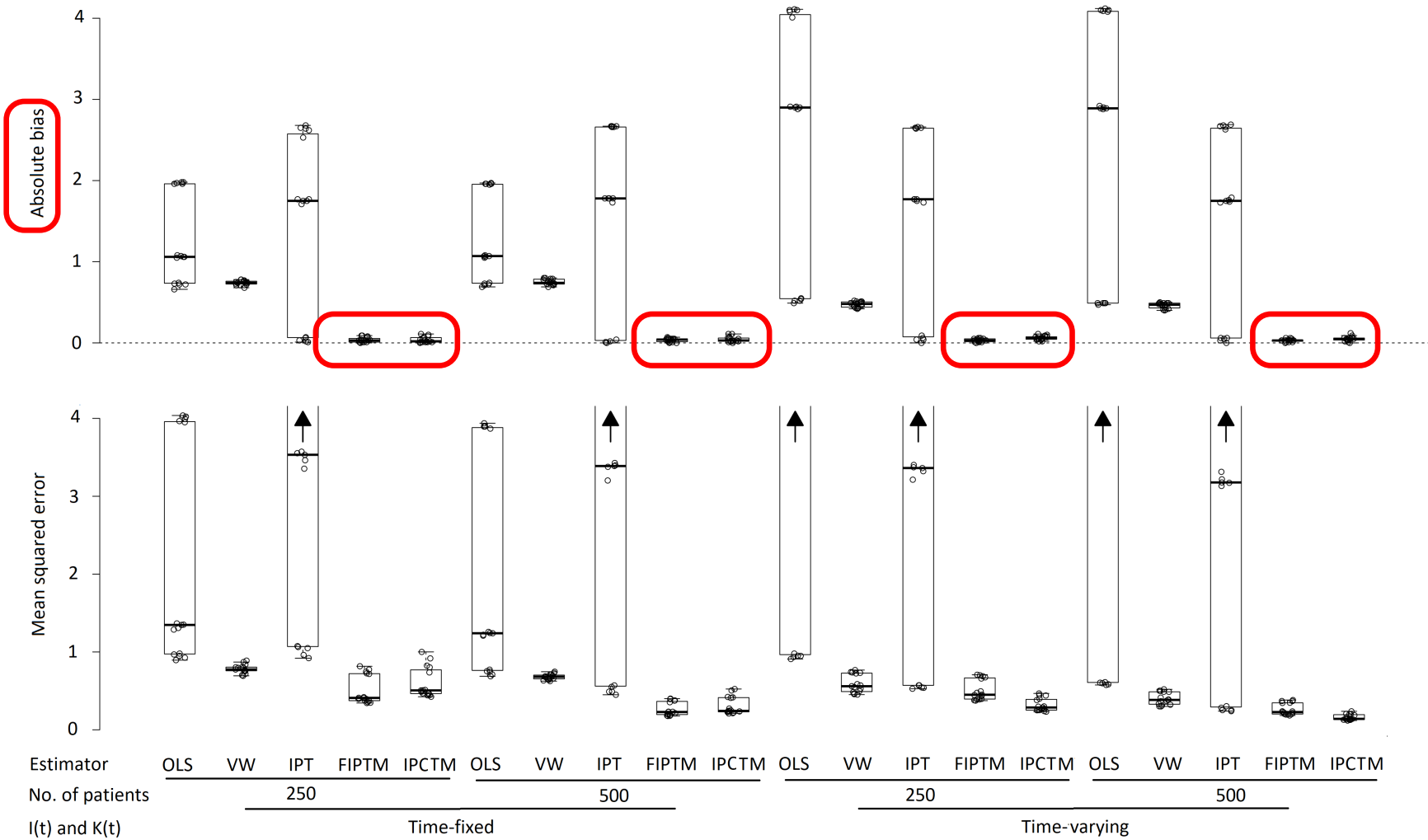
$$K_{1i}(0) \sim N(3,1), K_{1i}(t) = K_{1i}(t-1) + 0.01$$
$$K_{2i}(0) \sim Bern(0.55), K_{2i}(t) = K_{2i}(t-1)$$
$$K_{3i}(0) \sim N(-1.2,1), K_{3i}(t) \sim N(K_{3i}(t-1), 0.05)$$
$$I_i(t) \sim Bern(p_i(t))$$

with $p_i(t) = \dfrac{\exp(0.5 + 0.1K_{1i}(t) + 0.05K_{2i}(t) - 1K_{3i}(t) - 1.5I_i(t-1))}{1 + \exp(0.5 + 0.1K_{1i}(t) + 0.05K_{2i}(t) - 1K_{3i}(t) - 1.5I_i(t-1))}$

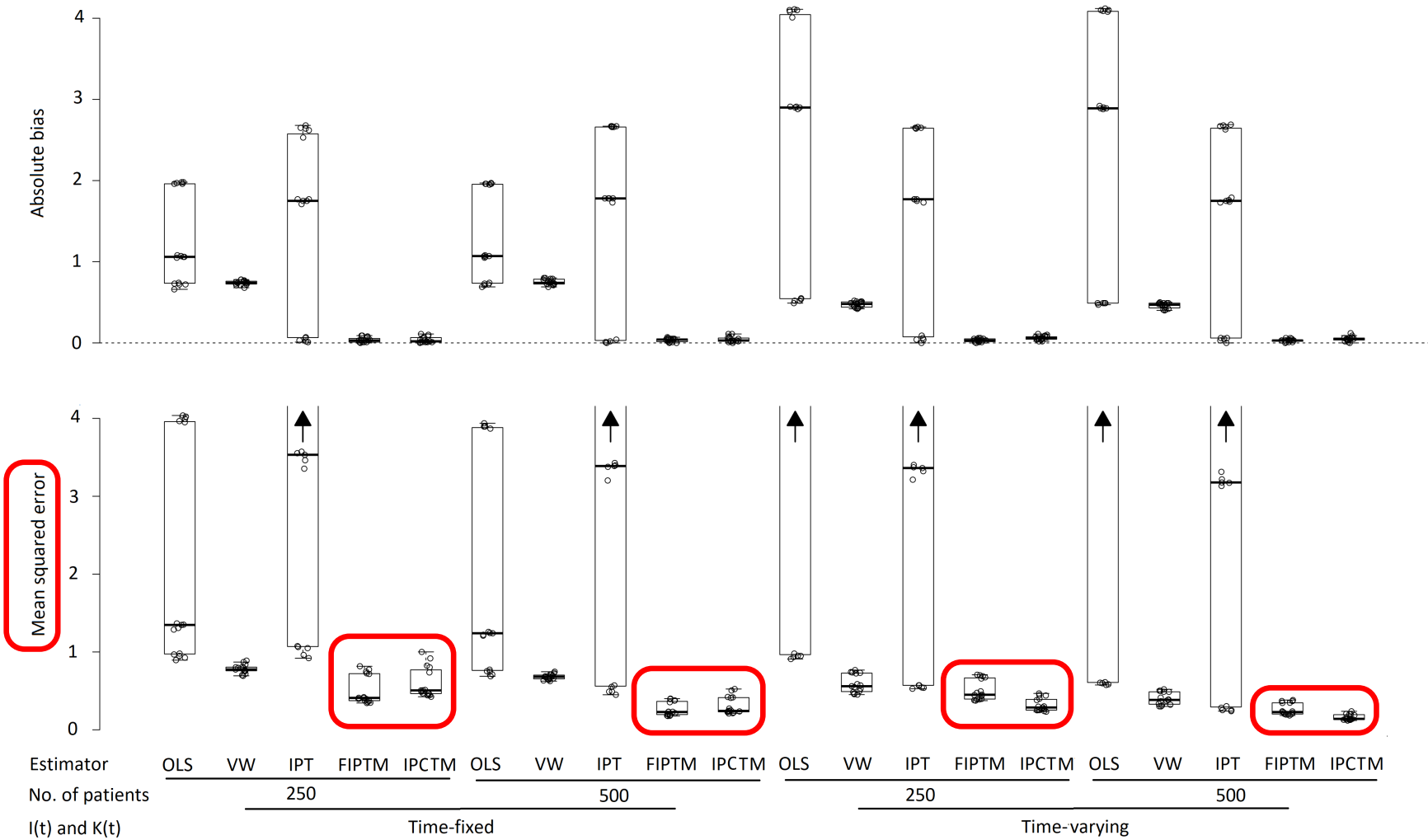
Results


Absolute bias
Mean squared error
15 γ and α(t) scenarios
Estimator
OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM
No. of patients
250
500
250
500
I(t) and K(t)
Time-fixed
Time-varying

Results


Absolute bias
Mean squared error
0
1
2
3
4
Estimator
OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM
No. of patients
250
500
250
500
I(t) and K(t)
Time-fixed
Time-varying

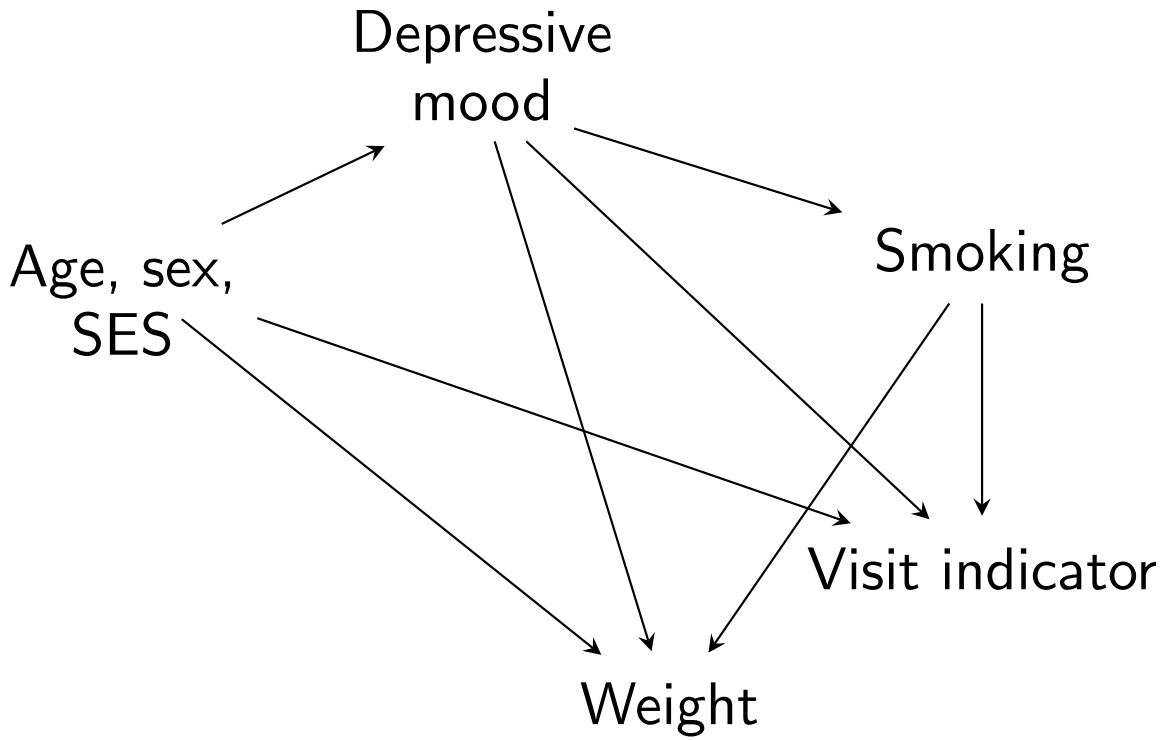
Results


Absolute bias
Mean squared error
Estimator
OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM OLS VW IPT FIPTM IPCTM
No. of patients
250
500
250
500
I(t) and K(t)
Time-fixed
Time-varying

Add Health analysis

Add Health study, Harris (2013)

- 4-wave longitudinal study of American adolescents followed until adulthood (1994-2008)
- Sample of 6504 adolescents at wave I
- Missing data (covariates and outcome) and/or dropout



Add Health study: Baseline characteristics

Table 1: Characteristics at baseline of children enrolled in the *Add Health* study, stratified by depressive mood

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

More smoking, greater age, more females and lower SES in adolescents with depressive mood at baseline.

Add Health study: Monitoring rate ratios

Table 2: Average rate ratios and 95% confidence intervals for variables in the proportional rate model for monitoring times

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

† Computed using Rubin's rule for multiply imputed datasets

Smoking, being younger and being a female were associated with greater chances for the weight to be recorded.

Add Health study: Marginal effect of depressive mood

Table 3: Comparison of the estimators of the marginal effect of depressive mood on weight in pounds

	Estimator	95% CI‡
$\hat{\beta}_{OLS}$	-3.83	-5.55; -2.11
$\hat{\beta}_{VW}$	-3.69	-5.44; -1.94
$\hat{\beta}_{FIPTM}$	1.43	-0.35; 3.21
$\hat{\beta}_{IPCTM}$	1.12	-0.59; 2.83

‡ Computed using bootstrapped variance and normal approximation

- Estimate is reversed when adjusting for both confounding and CDMT
- Effects consistent with the literature (Blaine, 2008; Van Strien, 2016)

Modelling monitoring

Considerations

Considerations for modelling the monitoring intensity:

- Which variables should we condition upon to block biasing paths? Are they measured?
- How to account for an history of covariates that varies by individual? (and in time!)
- What is the effect of time on the probability of visit at time t?

Visit intensity model

In the first project (effect of depressive mood), we assumed that the visits must only be modelled as a function of covariates $\mathbf{V}_i(t)$ and of time t to block any biasing path and postulated:

$$E[dN_i(t)|\mathbf{V}_i(t)] = \xi_i(t)\exp(\gamma'\mathbf{V}_i(t))d\Lambda_0(t).$$

Some authors extended this model and assumed that the covariates measured at the last visit affected the monitoring probability at time t. E.g., Zhu et al. (2017) assumed:

$$\mathbb{E}\left[dN_i(t)|\mathbf{V}_i(t-)\right] = \lambda_0(B_i(t))\exp(\gamma'\mathbf{V}_i(t-))dt$$

where $\mathbf{V}_i(t-)$ are observed at the last visit, and $B_i(t)$ is the time since the last visit, computed at time t.

Examples 1 and 2, Coulombe et al. (2021)

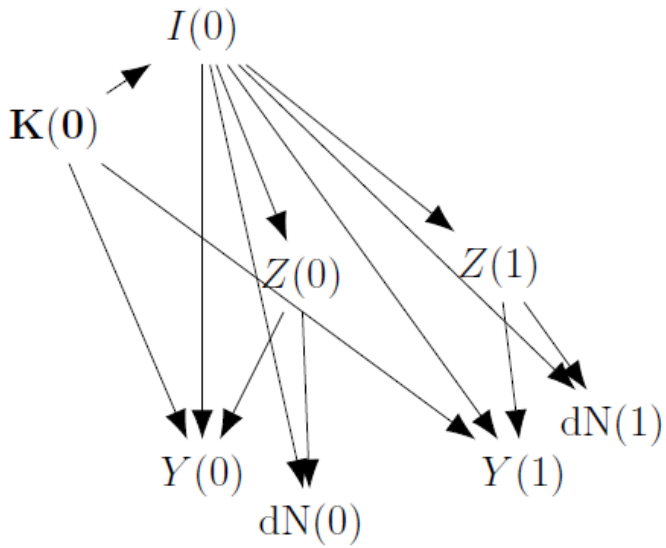


Figure 1: Causal diagram for the first DGM (patient index i removed) $I(0)$ is an intervention of interest whose marginal effect on a longitudinal outcome, $Y(t)$ – assumed to be time-invariant – is of interest. $\mathbf{K(0)}$ represent confounding variables, $Z(t)$ are mediators, and $dN(t)$ indicates the monitoring process through which the outcome is observed.

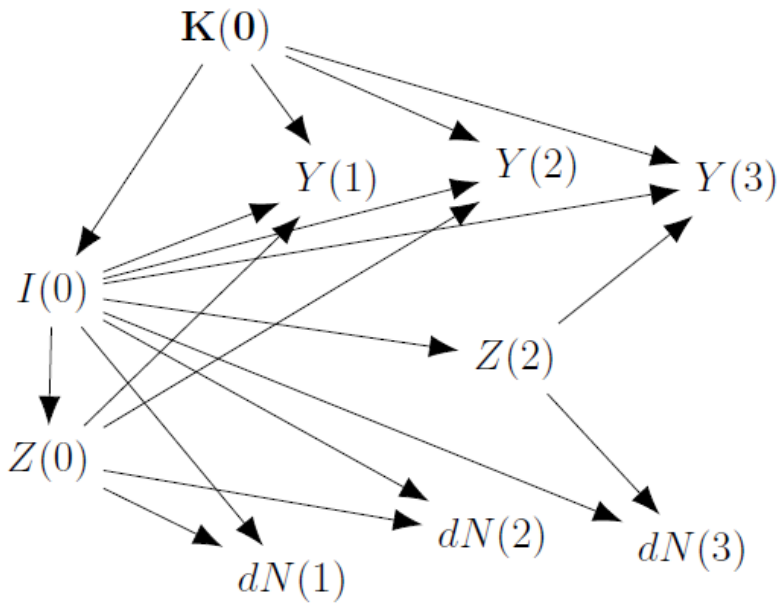


Figure 2: Causal diagram for the second DGM (patient index i removed) $I(0)$ is an intervention of interest whose marginal effect on a longitudinal outcome, $Y(t)$ – assumed to be time-invariant – is of interest. $\mathbf{K(0)}$ represent confounding variables, $Z(t)$ are mediators, and $dN(t)$ indicates the monitoring process through which the outcome is observed. Covariates $Z(t)$ are only "updated" at times 0 and 2 and affect next outcomes and monitoring indicators.

"Endogeneity"

Do standard inverse weights account for the following features:

- There is a cumulated effect of the time since last observation, and covariates "interact" or are modified by an observation time
- Having a physician visit can change health habits, nutritional habits, etc. Examples:
  - You visit your doctor, and she strongly suggests that you stop smoking or prescribes you a new medication
  - If I have a physician visit today, my chances of visiting tomorrow are smaller (gap time)
- All while evolving, patients' characteristics affect the outcome and the monitoring processes in time, creating long-term dependence between the two processes

Examples 3 and 4, Coulombe et al. (2021)

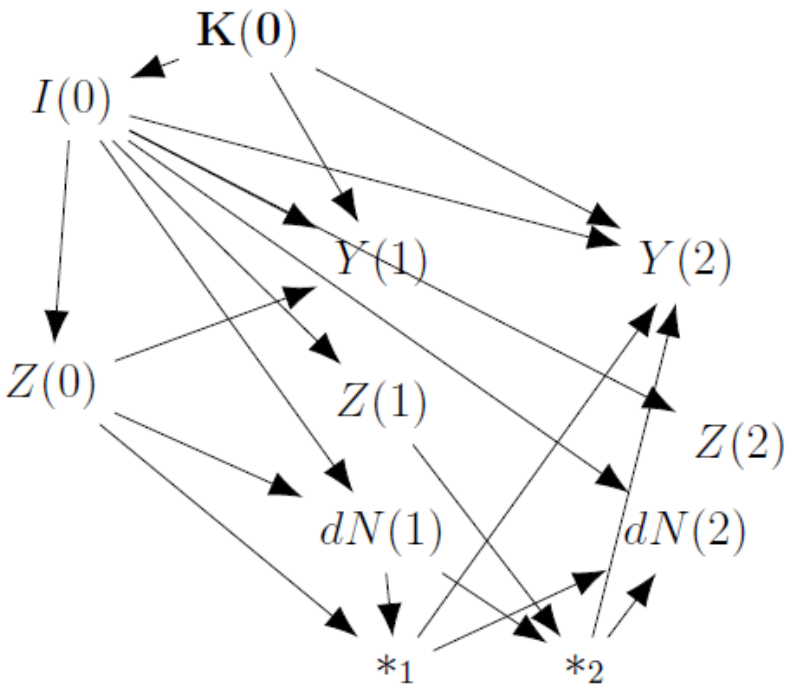


Figure 3: Causal diagram for the third DGM (patient index i removed) $I(0)$ is an intervention of interest whose marginal effect on a longitudinal outcome, $Y(t)$ – assumed to be time-invariant – is of interest. $\mathbf{K(0)}$ represent confounding variables, $Z(t)$ are mediators, and $dN(t)$ indicates the monitoring process through which the outcome is observed. Asterisks represent interactions between the covariates whose arrows point into it.

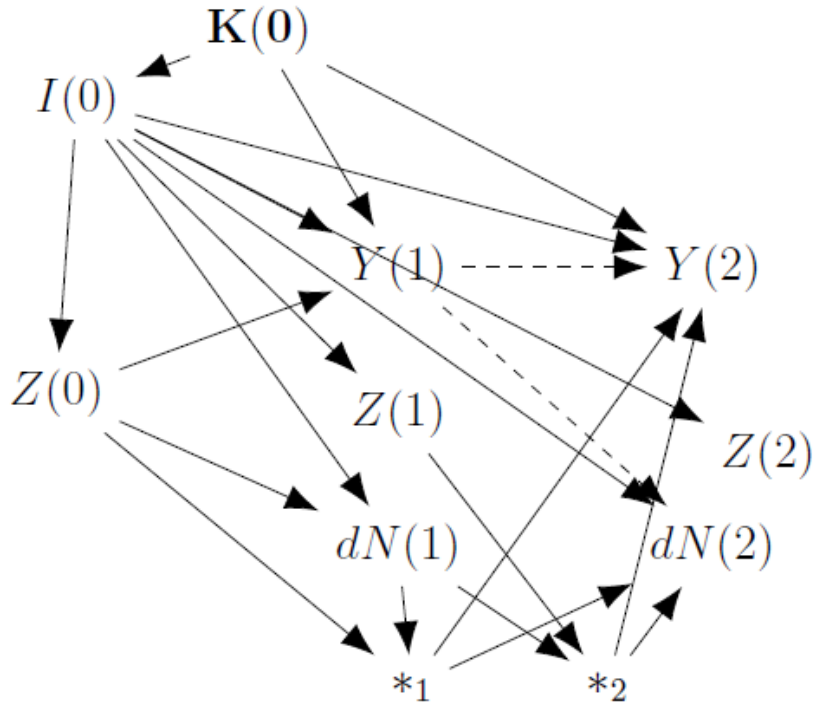


Figure 4: Causal diagram for the fourth DGM (patient index i removed) $I(0)$ is an intervention of interest whose marginal effect on a longitudinal outcome, $Y(t)$ – assumed to be time-invariant – is of interest. $\mathbf{K(0)}$ represent confounding variables, $Z(t)$ are mediators, and $dN(t)$ indicates the monitoring process through which the outcome is observed. Asterisks represent interactions between the covariates whose arrows point into it.

Proposed cumulated weight and stabilizers

For $p_i(t) = t - B_i(t)$ the last visit time and $B_i(t)$ the "gap time".

Proposed weight:

$$sw_{i,j}(t|\mathcal{H}^{\mathbf{o}}(\mathbf{t}-)) = \prod_{s=0}^{t} \left(\frac{\xi_i(s) \exp\left(\gamma_I I_i(0) + \boldsymbol{\gamma}_{\mathbf{Z}}' \mathbf{Z}_i(p_i(s))\right) \lambda_0(B_i(s))ds}{\lambda_{0,j}(B_i(s))ds} \right)^{\mathbb{I}(dN_i(s)=1)}$$
$$\times \left(\frac{1 - \xi_i(s) \exp\left(\gamma_I I_i(0) + \boldsymbol{\gamma}_{\mathbf{Z}}' \mathbf{Z}_i(p_i(s))\right) \lambda_0(B_i(s))ds}{1 - \lambda_{0,j}(B_i(s))ds} \right)^{\mathbb{I}(dN_i(s)=0)}$$

Proposed Breslow-type stabilizers:

$$\hat{\lambda}_{0,1}(B(t)) = \frac{\sum_{i=1}^{n} \int_{s=0}^{\tau} \mathbb{I}(dN_i(s) = 1 \cap B_i(s) = B(t))}{\sum_{i=1}^{n} \int_{s=0}^{\tau} \exp\left(\hat{\gamma}_I I_i(0) + \hat{\boldsymbol{\gamma}}_{\mathbf{Z}} \mathbf{Z}_i(p_i(s))\right) \mathbb{I}(dN_i(s) = 1 \cap B_i(s) = B(t))},$$

or $$\hat{\lambda}_{0,2}(B(t)) = \frac{\sum_{i=1}^{n} \int_{s=0}^{\tau} \mathbb{I}(dN_i(s) = 1 \cap B_i(s) = B(t))}{\sum_{i=1}^{n} \int_{s=0}^{\tau} \exp(\hat{\delta}_I I_i(0)) \mathbb{I}(dN_i(s) = 1 \cap B_i(s) = B(t))}.$$

Proposed cumulated weight and stabilizers cont'd

Solutions:

- No need to account for the full history of covariates because of the assumption on subsequent monitoring indicators
- Covariates in the nuisance models only need to be assessed at the same times as the outcomes
- First comparison of proposed stabilizers

Discussion

Discussion

Considerations:

- Other types of exposure and outcome (third doctoral project)
- Modelling depends on the problem you want to tackle (gap time, current time, lagged covariates, etc.)
- Other ways to model recurrent observation times: Andersen and Gill; Prentice, Williams and Peterson; Wei-Lin-Weissfeld model, etc.
- We rarely have a common observation schedule for other covariates (in nuisance models) and it's a challenge to get the necessary data to break the biasing paths (latent variables?)
- The methods discussed all rely on causal assumptions

Discussion cont'd

Areas of current/future work:

- Selection of variables in the visit intensity model
- Diagnostic tests (e.g., balance in characteristics across observed and non observed outcomes)
- Dynamic treatment regimes under irregular observation schedule
- Multiply robust estimators and semiparametric efficiency

Acknowledgments

Prof. Erica E. M. Moodie and Prof. Robert W. Platt

Prof. Marie Davidian and Prof. Shu Yang for their invitation and help with the organization

Alison McCoy for her help with the organization and advertisement

Feel free to be in touch!

janie.coulombe@mail.mcgill.ca

Funding and support from:

NSERC CRSNG

McGill UNIVERSITY

compute canada | calcul canada

Thank you

References

Andersen, P. K., and Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4), pp. 1100-1120.

Blaine, B. (2008) Does depression cause obesity? A meta-analysis of longitudinal studies of depression and weight control. *Journal of Health Psychology*, 13(8), pp. 1190-1197.

Buzkova, P., and Lumley, T. (2007) Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics*, 35(4), pp. 485-500.

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

Coulombe, J., Moodie, E. E. M., and Platt, R. W. (2021) Estimating the marginal effect of a continuous exposure on an ordinal outcome using data subject to covariate-driven treatment and visit processes. *Statistics in Medicine*, forthcoming.

Coulombe, J., Moodie, E. E. M., and Platt, R. W. (2021) Weighted regression analysis to correct for informative monitoring times and confounders in longitudinal studies. *Biometrics*, 77(1), pp. 162-174.

Greenland, S. (2003) Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3), pp. 300-306.

Harris, K. M. (2013) The Add Health study: Design and accomplishments. Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill, pp. 1-22.

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

Lin, H., Scharfstein, D. O., et Rosenheck, R. A. (2004) Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 66(3), pp. 791-813.

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

Neyman, J. S. (1923) On the application of probability theory to agricultural experiments. Essay on principles, section 9, *Statistical Science*, 5(14), pp. 465-472.

References cont'd

Pullenayegum, E. M., and Feldman, B. M. (2013) Doubly robust estimation, optimally truncated inverse-intensity weighting and increment-based methods for the analysis of irregularly observed longitudinal data. *Statistics in Medicine*, 12(6), pp. 1054-1072.

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

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

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

Sun, Y., McCulloch, C. E., Marr, K. A., et al. (2020). Recurrent events analysis with data collected at informative clinical visits in electronic health records. arXiv preprint arXiv:2004.12234.

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

Zero-mean counting process

Lin and Ying base their estimator on the process

$$M_i(t) = M_i(t; \beta, \gamma, \mathscr{A})$$
$$= \int_0^t \left[\left\{ Y_i(s) - \beta' X_i(s) \right\} dN_i(s) - \xi_i(s) \exp^{\gamma' Z_i(s)} d\mathscr{A}(s) \right].$$

Using assumptions $Y_i(s) \perp N_i(s) | V_i(s)$,
$E[dN_i(s)|V_i(s)] = \xi_i(s)\mathbb{E}[dN_i^*(s)|V_i(s)] = \xi_i(s) \exp^{\gamma' V_i(s)} d\Lambda_0(s)$,
and using the fact that $d\mathscr{A}(s) = d \int_0^t \alpha_0(s) dN_i(s) \Big|_{t=s} = \alpha_0(s) dN_i(s)$,
we can show it is zero-mean:

$$E\left[M_i(t)|X_i(t), Z_i(t)\right] =$$
$$E\left[\int_0^t \left[\left\{ Y_i(s) - \beta' X_i(s) \right\} dN_i(s) - \xi_i(s) \exp^{\gamma' Z_i(s)} d\mathscr{A}(s) \right] \Big| X_i(t), Z_i(t) \right]$$

Zero-mean counting process

$$
\begin{aligned}
&= E\left[\int_0^t \left(Y_i(s) - \beta' X_i(s)\right) dN_i(s) \middle| X_i(t), Z_i(t)\right] \\
&\qquad\qquad - E\left[\int_0^t \xi_i(s) \exp^{\gamma' Z_i(s)} d\mathscr{A}(s) \middle| X_i(t), Z_i(t)\right] \\
&= \int_0^t E\left[Y_i(s) dN_i(s) | X_i(t), Z_i(t)\right] - \int_0^t \beta_0' X_i(s) E\left[dN_i(s) | X_i(t), Z_i(t)\right] \\
&\qquad\qquad - \int_0^t \xi_i(s) \exp^{\gamma' Z_i(s)} E\left[d\mathscr{A}(s) | X_i(t), Z_i(t)\right] \\
&= \int_0^t E\left[Y_i(s) | X_i(t), Z_i(t)\right] E\left[dN_i(s) | X_i(t), Z_i(t)\right] \\
&\qquad\qquad - \int_0^t \beta_0' X_i(s) E\left[dN_i(s) | X_i(t), Z_i(t)\right] \\
&\qquad\qquad - \int_0^t \xi_i(s) \exp^{\gamma' Z_i(s)} \alpha_0(s) d\Delta_0(s)
\end{aligned}
$$

Zero-mean counting process

$$
\begin{aligned}
&= \int_0^t \left(\alpha_0(s) + \beta_0' X_i(s)\right) \xi_i(s) \exp^{\gamma' Z_i(s)} \, d\Delta_0(s) \\
&\qquad - \int_0^t \beta_0' X_i(s) \xi_i(s) \exp^{\gamma' Z_i(s)} \, d\Delta_0(s) \\
&\qquad\qquad - \int_0^t \alpha_0(s) \xi_i(s) \exp^{\gamma' Z_i(s)} \, d\Delta_0(s) \\
&= 0
\end{aligned}
$$

Simulation studies

N=1000, time discretized in 0.01-width units. Sample size and maximum follow-up time vary. Time-fixed confounders and exposure.

- Simulation study #1: Exogenous covariates
  - Marginal effect of covariates measured at time t on visit at time t
  - $P(dN_i(t) = 1|\mathbf{V_i(t)}) = \lambda_0(t)\exp(\gamma_V \mathbf{V_i(t)})$ with $I_i(t) \subset \mathbf{V_i(t)}$

- Simulation study #2: Endogenous covariates
  - Covariate process (that affects monitoring) is updated whenever there is a new visit
  - $P(dN_i(t) = 1|\mathbf{V_i}(T_i(t)), T_i(t)) = \lambda_0(t - T_i(t))\exp(\gamma_V \mathbf{V_i}(T_i(t)))$ with $I_i(0) \subset \mathbf{V_i}(T_i(t))$

  for $T_i(t)$ the last monitoring time in individual i, before time t.