Adaptive sparsening and smoothing of the treatment model for longitudinal causal inference using outcome-adaptive LASSO and marginal fused LASSO

Mireille Schnitzer[1,2,*], Denis Talbot[3,4], Yan Liu[1], David Berger[5], Guanbo Wang[6], Jennifer O'Loughlin[2,7], Marie-Pierre Sylvestre[2,7], and Ashkan Ertefaie[8]

[1]Faculté de pharmacie, Université de Montréal, Montréal, QC H3C 3J7, Canada
[2] Département de médecine sociale et préventive, Université de Montréal, Montréal, QC H3N 1X9, Canada
[3]Département de médecine sociale et préventive, Université Laval, Québec, QC G1V 0A6, Canada
[4]Centre de recherche du CHU de Québec, Université Laval, Québec, QC G1E 6W2, Canada
[5]Department of Computer Science, Université de Montréal, Montréal, QC H3T 1J4, Canada
[6]CAUSALab, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, United States
[7]Centre de recherche du centre hospitalier de l'Université de Montréal, Montréal, QC H2X 0A9, Canada
[8]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14627, United States

mireille.schnitzer@umontreal.ca

Abstract

Causal variable selection in time-varying treatment settings is challenging due to evolving confounding effects. Existing methods mainly focus on time-fixed exposures and are not directly applicable to time-varying scenarios. We propose a novel two-step procedure for variable selection when modeling the treatment probability at each time point. We first introduce a novel approach to longitudinal confounder selection using a Longitudinal Outcome Adaptive LASSO (LOAL) that will data-adaptively select covariates with theoretical justification of variance reduction of the estimator of the causal effect. We then propose an Adaptive Fused LASSO that can collapse treatment model parameters over time points with the goal of simplifying the models in order to improve the efficiency of the estimator while minimizing model misspecification bias compared with naive pooled logistic regression models. Our simulation studies highlight the need for and usefulness of the proposed approach in practice. We implemented our method on data from the Nicotine Dependence in Teens study to estimate the effect of the timing of alcohol initiation during adolescence on depressive symptoms in early adulthood.

Keywords: Causal Inference, LASSO, longitudinal data, marginal structural model, variable selection, inverse probability weighting.

1 Introduction

Near practical positivity violations are a common problem when conducting causal inference with time-varying binary treatments, especially when treatment can change over many time-points. For marginal structural models (MSMs), which are models for the counterfactual outcome under a longitudinal treatment intervention, the typical sequential positivity assumption requires that, at each time-point, the probability of accessing either level of treatment (propensity score) is non-zero for any possible covariate values and prior treatments. The related practical condition is that one must have observed study-end outcome information under all relevant treatment patterns and baseline and time-varying covariate values. In typical finite sample applications, this condition is difficult to satisfy [Rudolph et al., 2022, Schomaker et al., 2019]. Under such data sparsity, outcome regression-based methods can smooth over time points where treatments are not observed for certain covariate strata. However, unless all outcome models are correctly specified – which is difficult to achieve when extrapolation is necessary – this approach leads to biased estimation. Alternatively, many methods involve weighting by the inverse of the probability of treatment [Bang and Robins, 2005, Robins et al., 2000, van der Laan and Gruber, 2012]. This involves directly modeling the probability of treatment for each time-point. When smoothing is desirable, it is possible to pool these treatment models over time [Cole and Hernán, 2008, Hernán et al., 2000, Rudolph et al., 2022] or covariate information [Cole and Hernán, 2008], but this can also lead to bias if the resulting models are not correctly specified for the probability of treatment at each time-point. It is, therefore, of interest to develop data-adaptive approaches to model selection that can trade off bias and variance under sparse conditions.

A related challenge involves defining and selecting covariates to satisfy the sequential conditional exchangeability ("no unmeasured confounders") assumption [Cole and Hernán, 2008]. Beyond adjustment for confounders, excluding covariates that only affect the treatment (i.e. instruments) and adjusting for pure causes of the outcome can reduce estimation variance with inverse probability of treatment weighting [Brookhart et al., 2006, Rotnitzky et al., 2010, Schisterman et al., 2009]. Previous work related to fitting MSMs with time-varying treatments demonstrated quantitatively [Lefebvre et al., 2008] and theoretically [Rotnitzky and Smucler, 2020, Adenyo et al., 2024b] that estimation variance can be inflated when adjusting for variables that only cause treatment. In particular, Rotnitzky and Smucler [2020] established that, under a nonparametric model with a given directed acyclic graph (DAG), and a time-dependent adjustment set, the removal of certain types of non-confounding covariates will reduce the asymptotic variance of nonparametric efficient estimators.

Many approaches for covariate selection or reduction have been developed in a single time-point treatment setting [Loh and Vansteelandt, 2021, Persson et al., 2017, Schneeweiss et al., 2009, Tang et al., 2022], including Bayesian approaches [Talbot et al., 2015, Wang et al., 2012, Wilson and Reich, 2014]. In particular, the outcome-adaptive LASSO [Shortreed and Ertefaie, 2017] uses the inverse magnitude of the outcome model regression coefficients to penalize the corresponding covariates in an adaptive LASSO for the treatment model. This aims to exclude any variable that does not have a conditional association with the outcome. One version of Collaborative Targeted Maximum Likelihood Estimation (C-TMLE) [Gruber and van der Laan, 2010] greedily selects covariates into the treatment model when the inclusion improves the fit of the propensity score-updated outcome model, and uses cross-validation to select an optimal number of selection steps. Schnitzer et al. [2020] extended C-TMLE to the longitudinal treatment setting but noted its computational complexity.

While not yet applied to any causal setting, the fused LASSO [Höfling et al., 2010, Tibshirani et al., 2005, Viallon et al., 2016] was proposed to smooth over spacial or temporal structures by penalizing both coefficient magnitudes and the distance between coefficients of neighboring or

grouped covariates in a linear regression model. Viallon et al. [2013, 2016] proposed an extension for generalized linear models with adaptive weights [Zou, 2006] resulting in oracle properties. The optimization problem is solved with a coordinate-wise optimization algorithm [Höfling et al., 2010].

In this study, we explore a two-phase approach to treatment model dimension reduction in the longitudinal context with a causal objective of estimating the parameters of an MSM. We first define a "saturated" model for the probability of treatment at all time-points that adjusts for the complete history of covariates and completely stratifies the model by time. Then, we propose to remove covariates from this model using a longitudinal extension of the outcome-adaptive LASSO applied to the saturated model. Our selection criteria use parametric working outcome models to operationalize the variance reduction results in Rotnitzky and Smucler [2020] and Adenyo et al. [2024b]. The second step, carried out after the initial covariate reduction, involves model smoothing over time such that covariate-treatment associations can have shared parameters over time. This step utilizes an implementation of the adaptive fused LASSO for a logistic regression that penalizes the distance between coefficients of a given covariate and the treatment at different time points. We then perform extensive simulation studies to evaluate the performance and robustness of our approach compared to non-adaptive and oracle estimators and longitudinal C-TMLE. Finally, we apply our approach in a complex longitudinal setting to estimate the effect of drinking initiation in adolescents on scores of depression symptoms in early adulthood.

2 Data, target parameter, and estimation

In this section, we present the target of estimation and give preliminaries that will allow us to describe our proposed model selection procedure.

2.1 Data and target parameter

Suppose that we have a data structure $\boldsymbol{O} = (\boldsymbol{L}_0, A_0, \boldsymbol{L}_1, A_1, ..., A_T, Y)$ where the outcome Y is continuous, the treatments $A_t; t = 0, ..., T$ are binary, and the covariates $\boldsymbol{L}_t; t = 0, ..., T$ are multivariate. We use $\overline{\boldsymbol{L}}_t = (\boldsymbol{L}_0, ..., \boldsymbol{L}_t)$ to indicate the history of covariates up to and including time t. The covariates potentially include both binary and continuous components. We take n independent identically distributed samples, with realizations denoted by lowercase letters, e.g. $\boldsymbol{o}_i$ and $a_{0,i}$ are realizations of $\boldsymbol{O}$ and A_0, respectively, for $i = 1, ..., n$.

We consider an intervention that sets each treatment node A_t to some fixed value, either zero or one, i.e. setting $\overline{A} = (A_0, ..., A_T)$ to the treatment pattern $\overline{a} = (a_0, ..., a_T)$ where each a_t is either zero or one. Define the counterfactual variable $Y^{\overline{a}}$ to be the potential outcome under treatment pattern $\overline{a}$. Our interests lie in modeling the marginal expectation of the counterfactual outcome potentially conditional on some subset of the baseline covariates $\boldsymbol{L}_0$, i.e. modeling $\mathbb{E}(Y^{\overline{a}} \mid \boldsymbol{L}_0)$. As an example, define the MSM of interest as

$$\mathbb{E}(Y^{\overline{a}} \mid L_0^1) = \mu_0 + \mu_1 cum(\overline{a}) + \mu_2 L_0^1, \tag{1}$$

where $L_0^1 \subseteq \boldsymbol{L}_0$ is a single (one-dimensional) baseline covariate, and $cum(\overline{a})$ represents the cumulative function that counts the number of ones (i.e. number of treated time-points) in the treatment pattern $\overline{a}$. Thus, μ_1 represents the change in the expected outcome from one additional treated time point under the linear model and $\{\mu_0, \mu_1, \mu_2\}$ is our target parameter. We can more appropriately define the MSM as a working model, and the true parameter values as projections of the true counterfactual regression curve $\mathbb{E}(Y^{\overline{a}} \mid L_0^1)$ onto the working model [Neugebauer and van der Laan, 2007]. See Petersen et al. [2014] for complete details in the time-varying treatment context.

3

2.2 Identifiability

The MSM parameters are identifiable under typical causal assumptions including consistency, sequential positivity, and sequential conditional exchangeability. Sequential conditional exchangeability is given as

$$Y^{\overline{a}} \perp\!\!\!\perp A_t \mid (\overline{A}_{t-1} = \overline{a}_{t-1}, \overline{\boldsymbol{L}}_t), \qquad t = 0, ..., T,$$

where variables with negative subscripts should be discarded here and subsequently. Positivity means that $P(A_t \mid \overline{\boldsymbol{L}}_t, \overline{A}_{t-1}) > 0$ for all supported values of $(\overline{\boldsymbol{L}}_t, \overline{A}_{t-1})$. Consistency means that we can equate the potential outcome under the observed treatment to the observed outcome, i.e. $Y^{\overline{a}} = Y$ if $\overline{A} = \overline{a}$.

In order to review the identifiability of the MSM parameters, we define nested expectations of the outcome conditional on the treatment pattern of interest [Bang and Robins, 2005]. We initialize $q_{T+1}(\overline{a}_{T+1}, \overline{\boldsymbol{L}}_{T+1}) = Y$. We recursively define

$$q_t(\overline{a}_t, \overline{\boldsymbol{L}}_t) = \mathbb{E}\{q_{t+1}(\overline{a}_{t+1}, \overline{\boldsymbol{L}}_{t+1}) \mid \overline{A}_t = \overline{a}_t, \overline{\boldsymbol{L}}_t\}, \qquad t = T, ..., 0.$$

We will use the notation $q_t^{\overline{a}} = q_t(\overline{a}, \overline{\boldsymbol{L}}_t)$ for brevity. Under the above causal assumptions, the g-formula $\mathbb{E}(q_0^{\overline{a}} \mid L_0^1) = \mathbb{E}(Y^{\overline{a}} \mid L_0^1)$ identifies the true regression curve for $Y^{\overline{a}}$ with respect to $\overline{a}$ and L_0^1 [Bang and Robins, 2005]. Under the MSM in Equation (1), this also identifies the true values of the parameters $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2)$.

If we do not want to make any causal assumptions, we can alternatively define our parameter of inference statistically through the function $q_0^{\overline{a}}$. The parameters $\boldsymbol{\mu}$ can be directly defined as minimizing the least-squares risk function of the model with $q_0^{\overline{a}}$ as the outcome with a regression specification according to the right-hand side of Equation (1).

2.3 Estimators

There are many available estimators of $\boldsymbol{\mu}$. One such estimator is G-computation, which uses regressions to sequentially estimate each $q_t^{\overline{a}}$, starting at time T [Bang and Robins, 2005, Schnitzer et al., 2014]. Then, define q_t^s to be the vector composed of the stacked quantities $q_t^{\overline{a}}$ of each possible pattern $\overline{a}$. The final step involves regressing the estimate of the stacked vector q_0^s according to the MSM in Equation (1).

Inverse probability of treatment weighting (IPTW) is an alternative approach that involves estimating the functions $g_t(\overline{a}_t, \overline{\boldsymbol{L}}_t) = P(A_t = a_t \mid \overline{\boldsymbol{L}}_t, \overline{A}_{t-1} = \overline{a}_{t-1})$ for each time-point $t = 0, ..., T$. One implementation then involves regressing the observed Y on the covariates in Model (1) using weights equal to estimates of $w_t(\overline{a}_t, \overline{\boldsymbol{L}}_t) = \prod_{k=0}^t \frac{g_k(a_k, L_0^1)}{g_k(\overline{a}_k, \overline{\boldsymbol{L}}_k)}$ for $t = T$, where $g_t(a_t, L_0^1)$ is defined as the stabilizing probability $P(A_t = a_t \mid L_0^1)$. Longitudinal targeted maximum likelihood estimation (LTMLE) [Petersen et al., 2014, Schnitzer et al., 2014, Van der Laan and Gruber, 2012] is an approach that uses estimates of the functions $q_t^{\overline{a}}$ in addition to the weights $w_t(\overline{a}_t, \overline{\boldsymbol{L}}_t)$ in order to estimate the parameters of the MSM.

A primary question across all methods that use inverse probability of treatment weights (such as IPTW and LTMLE) is how to approach modeling the treatment process to estimate the functions $g_t(\overline{a}_t, \overline{\boldsymbol{L}}_t)$. The two primary approaches are to model the treatment separately at each time-point or to pool the model over the $T + 1$ time-points. The latter is interesting because it allows for model simplification under sparsity while still allowing for greater model complexity with sufficient data support. But without a priori restrictions on the conditioning of the pooled treatment model, it is clear that the number of covariates can become large as the number of time-points increases. And

indeed, incorrect pooling and modeling decisions can lead to bias in the estimation of the MSM parameters.

3 Variable Selection

In this section, we describe covariate reduction of the treatment model using an implementation of the outcome-adaptive LASSO.

3.1 Selection goal

We consider parametric logistic regressions for the treatment models. For simplified illustration, we consider two time-points ($T = 1$), with the associated data structure $O = (\boldsymbol{L}_0, A_0, \boldsymbol{L}_1, A_1, Y)$. We consider a model for the probability of treatment, stratified on time, written as

$$\text{logit}\left\{P(A_0 = 1 \mid \boldsymbol{L}_0)\right\} = \alpha_{0,-1} + \boldsymbol{\alpha}_{0,0}\boldsymbol{L}_0, \tag{2}$$

$$\text{logit}\left\{P(A_1 = 1 \mid \overline{\boldsymbol{L}}_1, A_0)\right\} = \alpha_{1,-1} + \boldsymbol{\alpha}_{1,0}\boldsymbol{L}_0 + \boldsymbol{\alpha}_{1,1}\boldsymbol{L}_1 + \alpha_{1,2}A_0, \tag{3}$$

where the coefficients may be vectors when the corresponding $\boldsymbol{L}_t$ is multivariate. We can represent the same restriction on the mean as the pooled model

$$
\begin{aligned}
m_t(\overline{\boldsymbol{L}}_t, \overline{A}_{t-1}; \boldsymbol{\alpha}) =& \text{logit}\left\{P(A_t = 1 \mid \overline{A}_{t-1}, \overline{\boldsymbol{L}}_t)\right\} = \mathbb{I}(t = 0)\left(\alpha_{0,-1} + \boldsymbol{\alpha}_{0,0}\boldsymbol{L}_0\right) \\
&+ \mathbb{I}(t = 1)\left(\alpha_{1,-1} + \boldsymbol{\alpha}_{1,0}\boldsymbol{L}_0 + \boldsymbol{\alpha}_{1,1}\boldsymbol{L}_1 + \alpha_{1,2}A_0\right),
\end{aligned}
\tag{4}
$$

where $t \in \{0, 1\}$ and $\mathbb{I}(\cdot)$ is the indicator function.

Define $\boldsymbol{\alpha} = (\alpha_{0,-1}, ..., \alpha_{1,2})$ which are the coefficients of the covariates in the pooled propensity score model in (4).

Now consider the working regression models

$$E(q_0^{\overline{a}} \mid \boldsymbol{L}_0) = \beta_{0,-1} + \boldsymbol{\beta}_{0,0}\boldsymbol{L}_0, \tag{5}$$

$$E(q_1^{\overline{a}} \mid \overline{\boldsymbol{L}}_1) = \beta_{1,-1} + \boldsymbol{\beta}_{1,0}\boldsymbol{L}_0 + \boldsymbol{\beta}_{1,1}\boldsymbol{L}_1 + \beta_{1,2}a_0, \tag{6}$$

with true parameter values minimizing the risk under a squared-error loss function. Note that under the causal assumptions, these correspond to working structural models for $Y^{\overline{a}}$, i.e. Model (5) for $E(Y^{\overline{a}} \mid \boldsymbol{L}_0)$ and Model (6) for $E(Y^{\overline{a}} \mid \overline{\boldsymbol{L}}_1)$. Denote $\boldsymbol{\beta} = (\beta_{0,-1}, ..., \beta_{1,2})$ and let $\boldsymbol{\beta}^\dagger = (\beta_{0,-1}^\dagger, ..., \beta_{1,2}^\dagger)$ be an indicator vector of the non-zero elements of $\boldsymbol{\beta}$, fixing the intercept and treatment terms as non-zero, i.e. $\beta_{0,-1}^\dagger = \beta_{1,-1}^\dagger = \beta_{1,2}^\dagger = 1$.

We characterize the specific objectives of our variable selection as:

Objective 1 *For each time-point, select variables into the treatment model at time t that have corresponding non-zero coefficients $\boldsymbol{\beta}$ in the model for $q_t^{\overline{a}}$. We estimate the coefficients of the propensity scores*

$$m_t(\overline{\boldsymbol{L}}_t, \overline{A}_{t-1}; \boldsymbol{\alpha}^\dagger), \quad t = 0, 1, \tag{7}$$

where $\boldsymbol{\alpha}^\dagger$, of the same length as $\boldsymbol{\alpha}$, has fixed elements equal to zero corresponding to the zero items in $\boldsymbol{\beta}^\dagger$. The optimal value of $\boldsymbol{\alpha}^\dagger$, denoted as $\boldsymbol{\alpha}_0^\dagger$ (with the same elements fixed at zero), minimizes the risk under the logistic quasi-log-likelihood loss function.

5

This specific variable selection criterion can be motivated as removing covariates from the adjustment set $\overline{\boldsymbol{L}}_T$ that are not associated with the potential outcome $Y^{\overline{a}}$ conditional on the remaining variables in $\overline{\boldsymbol{L}}_T$, and the past treatment $\overline{A}_{T-1}$. In particular, these covariates are not relevant for sequential conditional exchangeability (i.e. are not confounders). However, the criterion retains variables that are conditionally associated with the potential outcome, regardless of whether they are confounders. This is an operationalization of the identification of a covariate subset that, when removed from the adjustment set, leads to a reduction in estimation variance in the nonparametric model [Rotnitzky and Smucler, 2020, Adenyo et al., 2024b]. Objective 1 also corresponds with the recommendations in Lefebvre et al. [2008] to only adjust for variables that affect the outcome through pathways that do not include treatment.

3.2 Longitudinal Outcome Adaptive Lasso (LOAL)

In order to write out the estimator, we expand the notation of the possibly multivariate covariates. First, we use $\tau = 0, 1$ to index the propensity score model for A_τ (i.e. Models (2) and (3)). We use $t = 0, 1$ to index the covariates $\boldsymbol{L}_t$ as before, where $\boldsymbol{L}_0 \in \mathbb{R}^{p_0}$ and $\boldsymbol{L}_1 \in \mathbb{R}^{p_1}$ such that p_t is the number of covariates of $\boldsymbol{L}_t$. The kth component of $\boldsymbol{L}_t$ is denoted $L_{t,k}; k = 1, \ldots, p_t$. Denote the set $\mathcal{J} = \{(0, 1, \mathcal{J}_{0,0}), (1, 1, \mathcal{J}_{1,0}), (1, 2, \mathcal{J}_{1,1})\}$ as the 3-dimensional indices of the coefficients $\boldsymbol{\alpha}$ being shrunk, where the set $\mathcal{J}_{\tau,t}$ indexes the specific covariates in $\boldsymbol{L}_t$ being shrunk within propensity score model τ. Note that overlapping indices in $\mathcal{J}_{0,0}$ and $\mathcal{J}_{1,0}$ index coefficients for the same covariates in different models. For example, $(0, 0, 1)$ and $(1, 0, 1)$ refer to the coefficients of covariate $L_{0,1}$ in Models (2) and (3), respectively, or the equivalent in Model (4). Also note that the intercept coefficients (indices $(0,-1)$ and $(1,-1)$ in Model (4)), as well as the coefficients associated with treatment (index $(1,2)$ in Model (4)) are not candidates for shrinking and so are excluded from $\mathcal{J}$. The indices in $\mathcal{J}$ are similarly used to refer to the corresponding coefficients $\boldsymbol{\beta}$ in Models (5) and (6).

Suppose that we have estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in Models (5) and (6) that are $\sqrt{n}$-consistent where n is the sample size. Given a regularization parameter $\lambda_n \geq 0$, an outcome-adaptive LASSO estimator of $\boldsymbol{\alpha}^\dagger$ in the pooled Model (4) as defined in Objective 1 is given as

$$\hat{\boldsymbol{\alpha}}(\lambda_n) = \arg\min_{\alpha} \sum_{\tau=0}^{1} \sum_{i=1}^{n} \left[a_{\tau,i} \log\{m_\tau(\overline{\boldsymbol{l}}_{\tau,i}, \overline{a}_{\tau-1,i}; \boldsymbol{\alpha})\} \right.$$
$$\left. + (1 - a_{\tau,i}) \log\{1 - m_\tau(\overline{\boldsymbol{l}}_{\tau,i}, \overline{a}_{\tau-1,i}; \boldsymbol{\alpha})\} \right] + \lambda_n \sum_{j \in \mathcal{J}} \hat{\omega}_j |\alpha_j|,$$

where $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma}$ for all $j \in \mathcal{J}$, with tuning parameter $\gamma > 1$. By the results in Shortreed and Ertefaie [2017], this estimator is asymptotically normal and consistent for the selection of covariates in the Model (7) if we assume that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{\gamma/2-1} \to \infty$. Note that $\gamma > 2$ is needed for the second convergence requirement.

For implementation purposes, this regularized regression can be run using a transformation of the pooled data, setting $V_{0,-1} = 1 - \mathbb{I}(t = 1)$, $\boldsymbol{V}_{0,0} = \boldsymbol{L}_0 - \mathbb{I}(t = 1)\boldsymbol{L}_0$, $V_{1,-1} = \mathbb{I}(t = 1)$, $\boldsymbol{V}_{1,0} = \mathbb{I}(t = 1)\boldsymbol{L}_0$, $\boldsymbol{V}_{1,1} = \mathbb{I}(t = 1)\boldsymbol{L}_1$, and $V_{1,2} = \mathbb{I}(t = 1)A_0$ with respectively corresponding coefficients $\alpha_{0,-1}, \ldots, \alpha_{1,2}$ in Model (4). Then, the adaptive LASSO is run with pooled outcome A_t on covariates $V_{0,-1}, \ldots, V_{1,2}$, without an intercept term, using weights $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma}; \forall j \in \mathcal{J}$.

3.3 Estimation of $q_t^{\overline{a}}$ to estimate $\boldsymbol{\beta}$

The proposed variable selection for the propensity scores is based on the estimated β parameters in Models (5) and (6), which need to be estimated at $\sqrt{n}$ rates. However, this requires preliminary

estimates of $q_1^{\bar{a}}$ and $q_0^{\bar{a}}$. To get these, we could first use a flexible regression method to estimate $q_1^{\bar{a}}$ by regressing Y on $\overline{L}_1$ and $\overline{A}_1$. We then generate predictions from this model for each pattern of interest $\bar{a}_1 = (a_0, a_1)$. We then run a pooled regression of the stacked vector q_1^s on the covariates $\overline{L}_1$ and a_0 where a_0 takes the value zero or one depending on the pattern $\bar{a}_1$, corresponding to the working structural Model (6). This results in estimates of the coefficients in that model, denoted by $\hat{\beta}_{1,-1}, ..., \hat{\beta}_{1,2}$.

For each pattern $\bar{a} \in \mathcal{A}$ where $\mathcal{A}$ is the set of all possible patterns, use a flexible regression method to regress $q_1^{\bar{a}}$ on $\boldsymbol{L}_0$ and A_0. We then use this model to make predictions setting $A_0 = a_0$ to obtain $q_0^{\bar{a}} = \mathbb{E}(q_1^{\bar{a}} \mid \boldsymbol{L}_0, a_0)$. We then run a pooled regression of the stacked vector q_0^s on the covariate $\boldsymbol{L}_0$ according to the structural Model (5) to obtain estimates of $\beta_{0,-1}$ and $\beta_{0,0}$, which we will denote $\hat{\beta}_{0,-1}$ and $\hat{\beta}_{0,0}$, respectively.

3.4 Selection of the tuning parameters

To select values for the two tuning parameters of the LOAL, we first fix the value of γ to a value slightly larger than 2 (we used 2.5 in the simulation and application) in order to ensure the required divergence of $\lambda_n n^{\gamma/2-1}$. We propose to select λ_n using a one-dimensional extension of the weighted absolute mean difference proposed in Shortreed and Ertefaie [2017], corresponding to a summary of a longitudinal balancing metric over covariates and times. See Web Appendix A for details.

4 Selective Fusion

In this section, we describe a second approach to dimension reduction which adaptively pools related coefficients across the time-point-specific treatment models using the fused LASSO.

4.1 Fusion goal

The Model (7) may not be sufficiently parsimonious in the following situation: suppose that, for some k, $\beta_{0,0,k}$ and $\beta_{1,0,k}$, the coefficients of covariate $L_{0,k}$ in the two structural models, are both large. This will mean that little penalty will be placed on the coefficients $\alpha_{0,0,k}$ and $\alpha_{1,0,k}$ in the LOAL procedure. In the situation where, for some $k \in \mathcal{J}_{0,0} \cap \mathcal{J}_{1,0}$, there is little difference in log-odds between $L_{0,k}$ and A_1 relative to $L_{0,k}$ and A_0, conditional on other terms in the model, we want the LOAL to fuse the terms $\alpha_{0,0,k}$ and $\alpha_{1,0,k}$. That is, we want to set $\alpha_{0,0,k} = \alpha_{1,0,k}$ or equivalently, have a single time-independent coefficient for $L_{0,k}$. This has the effect of smoothing over time and is the finite-sample Objective 2.

Objective 2 *In finite samples, fuse coefficients for common covariates across treatment models at different time points if it improves the pooled treatment model fit.*

By reducing the number of degrees of freedom, and avoiding potential overfitting of the propensity score models, we expect that this objective will lead to more efficient estimation of $\boldsymbol{\mu}$ in finite samples and avoid non-data-driven smoothing decisions under data sparsity.

4.2 Estimation

In order to achieve Objective 2, we first obtain the estimates $\hat{\boldsymbol{\alpha}}^{\text{refit}}(\lambda_n)$ from the LOAL and a refitted logistic regression, and define $\boldsymbol{\alpha}^*$ as the parameter vector of the same length as $\boldsymbol{\alpha}$ that is set to zero

at the indices of the zero-elements of $\hat{\boldsymbol{\alpha}}^{\text{refit}}(\lambda_n)$. Then we use a generalized Adaptive Fused LASSO [Viallon et al., 2016]

$$\underset{\boldsymbol{\alpha}^*}{\arg\min} \sum_{\tau=0}^{1} \sum_{i=1}^{n} \left[a_{\tau,i} \log\{m_\tau(\bar{l}_{1,i}, a_{0,i}; \boldsymbol{\alpha}^*)\} + (1 - a_{\tau,i}) \log\{1 - m_\tau(\bar{l}_{1,i}, a_{0,i}; \boldsymbol{\alpha}^*)\} \right]$$

$$+ \lambda_{1,n} \sum_{k \in \mathcal{J}_{0,0}^* \cap \mathcal{J}_{1,0}^*,} \frac{|\alpha_{1,0,k}^* - \alpha_{0,0,k}^*|}{|\hat{\alpha}_{1,0,k}^{\text{refit}}(\lambda_n) - \hat{\alpha}_{0,0,k}^{\text{refit}}(\lambda_n)|^{\gamma_1}},$$

with $\gamma_1 > 0$ and where $\mathcal{J}_{0,0}^* \subset \mathcal{J}_{0,0}$ and $\mathcal{J}_{1,0}^* \subset \mathcal{J}_{1,0}$ represent the indices of the selected covariates at $\tau = 0$ and 1, respectively. We propose to select the tuning parameters γ_1 and $\lambda_{1,n}$ by Bayesian information criterion (BIC) [Viallon et al., 2016]. We omit a sparsity-inducing penalty for two reasons: our variable selection was performed in the separate first step, and the variable selection and fusion objectives have different statistical goals (covariate balance vs. model selection, respectively).

It is important to note that, due to non-collapsibility and collinearity, the values of the coefficients in the pooled treatment models depend on the other covariates in the model. So, we first need to identify the covariate set before being able to statistically determine whether or not two coefficients should be fused. Thus, our application of the adaptive fused LASSO with a logistic regression model involves a purposeful misspecification of the pooled treatment model where we have already potentially marginalized over covariates in the previous step. We define a graph over the remaining covariates indicating which we allow to fuse. The oracle results of Viallon et al. [2016] apply relative to the model marginalized over the covariates removed in the first step, assuming that the treatment follows a Bernoulli distribution with a mean that is logit-linear in the remaining covariates. We also need that $\hat{\boldsymbol{\alpha}}^{\text{refit}}(\lambda_n)$ converges to $\boldsymbol{\alpha}^\dagger$ at a $\sqrt{n}$-rate, which is supported by the LOAL theory. Our application of the adaptive fused LASSO does not include a sparsity component in the objective function (i.e. does not include a LASSO penalty for the sizes of the coefficients). Through the same arguments, the oracle results then hold for the fusion of equal parameters α_k that are connected in the graph. A formal statement is given in Web Appendix B.

4.3 Fusion with $T > 1$

This procedure can be expanded when the number of time-points is greater than 2. The Fused LASSO requires the user to define a graph indicating which coefficients are allowed to fuse [Viallon et al., 2016]. This graph should represent the maximum smoothing of the model through data pooling, corresponding perhaps to how propensity score models are typically pooled. For baseline covariates, a "clique graph" may be used where we allow the coefficients of common variables to fuse between times $\tau = 0, ..., T$. Alternatively, we may use a "chain graph" where each chain connects the coefficient of the same baseline variable at successive time points. For common time-updated covariates, one may fuse coefficients of variables with the same lag relative to the treatment time τ across times τ, for which we could use a chain graph for successive fusing or a clique graph for fusing between any two time-points. We illustrate the usage of clique graphs with lagged time-dependent variables in the application.

5 Simulations

In this simulation study, we estimated the coefficients in the MSM (1) where L_0^1 was always defined as the first confounder in the dataset. We applied LOAL and the two-step fused LOAL to estimate

the three target parameters in the MSM (1). For benchmarks, we also ran the sequential G-computation [Bang and Robins, 2005] with all covariate main terms, IPTW with treatment models stratified by time-point and including all covariate main terms ("full IPTW"), and IPTW with pooled treatment models excluding all unwanted covariates ("IPTW oracle select") and further with correctly fused coefficients for common terms ("IPTW oracle select and fuse"). For fair comparisons, the specification of the models for estimating each q_t was common across methods that used these quantities.

The LOAL was implemented using adaptive weights in `glmnet` [Friedman et al., 2010]. We set $\gamma = 2.5$ (which allows for the convergence of $\lambda_n n^{\gamma/2-1}$) and a very broad range for candidate λ values and then selected the optimal λ_n value according to the balance criterion (Web Appendix A). We performed the fusion step using the archived `FusedLasso` package [Tibshirani, 2011] which implements the coordinate-wise optimization algorithm of Höfling et al. [2010], implementing adaptive weights for fusion as in Viallon et al. [2016] but setting the adaptive weights for the main terms to zero so that no additional sparsity would be induced. We set $\gamma_1 = 2.5$ and searched over a very broad range for λ_1, selecting the optimal $\lambda_{1,n}$ using the BIC.

5.1 Scenario 1: low dimensional with two time-points

In this scenario, we generated i.i.d. data $O = (C_0, I_0, A_0, C_1, I_1, A_1, Y)$ according to the left DAG in Figure 2, where C_0, and I_0 were independently generated from a standard normal distribution and A_0 and A_1 were Bernoulli distributed. Variables C_1 and I_1 were Gaussian-distributed with means $(A_0 + C_0)$ and (C_0), respectively. The instruments I_0 and I_1 only affected the treatment probabilities. Notably, A_0 was equal to one with probability $\text{logit}^{-1}(1.515C_0 + I_0)$ and A_1 with probability $\text{logit}^{-1}(-0.5 + 0.5C_0 + 0.25C_1 + 0.5A_0 + I_1)$. This made it so that in the marginal treatment models without instruments, the coefficients of C_0 in each model were both equal to 1.28. No other coefficients were equal. All four covariates were standardized to zero mean and unit standard deviation.
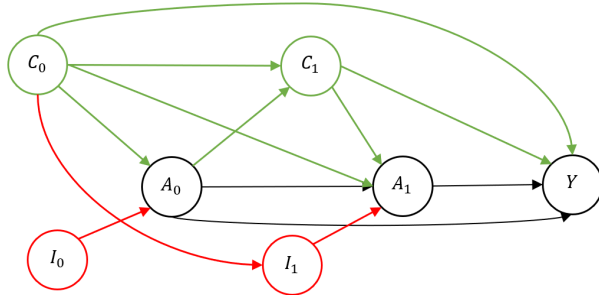

Figure 1: DAG representing the data generation in Scenarios 1. The target variable selection retained all covariates labeled C.

In order to evaluate the robustness of our method to misspecification of the working structural models, the Gaussian-distributed outcome Y had a mean specified in three ways: a) as a function of the main terms C_0, A_0, C_1, and A_1, b) the same as (a) but with an added interaction between C_0 and C_1 (covariate interaction) and c) the same as (a) but with an added interaction between A_0 and C_1^2 (effect modification). We specified the linear model for q_1 as containing only the main terms of all covariates; the linear model for q_0 also contained interactions between C_0 and A_0 and a squared term for I_0. The complete data generating mechanism is given in Web Appendix C.1.

We varied the sample sizes from $n = 200, 500, 1000$. Table 1 reports $\sqrt{n}$ times the absolute bias

9

and n times Monte Carlo mean squared error (MSE) over 1,000 draws for each estimator for these three outcome generating models. The variable selection results for the proposed estimators are given in Table 2. In Scenario 1(a) where the models used for the outcome process were correctly specified, the G-computation estimator was unbiased with the lowest standard errors. All IPTW estimators were unbiased since they adjusted for all confounders; the oracle IPTWs had lower MSE than the IPTW adjusting for all covariates ("full IPTW") though there was no difference between the two oracles. LOAL performed as well as the oracles, as did the fused LOAL. In Scenario 1(b), all outcome models were misspecified and G-Computation was the most biased. The full and oracle IPTW estimators were consistent but held some bias at these sample sizes; the oracle IPTW estimators had similar MSEs that were smaller than the full IPTW. LOAL and fused LOAL had lower MSE than the oracle IPTWs and comparable bias. In Scenario 1(c), the G-computation was highly biased but the full and oracle IPTW estimators were consistent with little bias. The oracles had the lowest MSE. The LOAL and fused LOAL were more biased than the oracle IPTWs and had higher MSE, but were still less biased than the G-computation and had lower MSE than the full IPTW.

In Table 2, we give the proportion of each selected variable and fused coefficient for each of the proposed methods. At the top of the table, we see that, when the oracle estimates of α (i.e. estimated with correctly selected propensity score models) are used to weight the fused LASSO, the L_0 coefficients correctly fused 99% of the time at all sample sizes. In Scenario 1(a), LOAL correctly selected the covariates with non-zero β coefficients in the working structural models $79 - 100\%$ of the time for all sample sizes, with greater true positive rates for larger β values. The method also correctly omitted the instruments with almost no false positives by $n = 500$. The success of fusion notably depended on the success of the variable selection, so that the L_0 covariates fused when the true model was selected in the first phase. In Scenarios 1(b) and 1(c) where the models to estimate q_t were misspecified, the convergence of the covariate selection was slower. In Scenario (c), the $\hat{\beta}$ did not converge to zero for the instruments, making it so that de-selection of instruments was not consistent (though the selection of confounders was consistent, but with slow convergence). Since fused LASSO relies on the correct covariate selection, the proportion of fusion was lower than for the other scenarios.

5.2 Additional simulation scenarios and results

We additionally ran a higher dimensional scenario which demonstrated the good performance of the LOAL and Fused LOAL in terms of estimation, variable selection, and fusion, with 30 covariates and two time points. We ran a third scenario with five time-points, demonstrating primarily that the smoothing by Fused LOAL approaches the oracle estimation and can positively impact estimation variance. Finally, we compared the performance of LTMLE, implemented with and without LOAL, and C-LTMLE (a comparator variable selection method) in the first two scenarios. While C-LTMLE is slightly better in terms of MSE when the outcome models are correctly specified, LTMLE with LOAL performs better when they are not. The major benefit of LOAL over C-LTMLE is dramatically lower computational complexity. See Web Appendix C.2-4 for details.

6   Example: The Nicotine Dependence in Teens (NDIT) study

We now illustrate the application of our proposed methodology using data from the Nicotine Dependence in Teens (NDIT) study to estimate the effect of the timing of alcohol initiation during adolescence on depressive symptoms in early adulthood. The NDIT study is a prospective longitudinal study initiated in 1999-2000, comprising 1,294 grade seven students recruited from 10 high

Method\Scenario	a) Outcome model with main terms			b) Outcome model with covariate interaction			c) Outcome model with effect modification		
	μ_0 -1.5	μ_1 1.5	μ_2 1.25	μ_0 1	μ_1 2.75	μ_2 1.25	μ_0 -1.5	μ_1 4	μ_2 5
True values									
n=200									
G-comp main terms	0.0(37)	0.1(24)	0.0(21)	7.9(183)	0.2(185)	0.9(110)	29.4(461)	20.2(362)	12.0(232)
IPTW full main terms	0.3(55)	0.4(43)	0.2(40)	5.7(224)	0.1(261)	0.2(145)	1.9(256)	1.8(330)	3.4(188)
IPTW oracle select	0.2(46)	0.3(36)	0.2(33)	4.7(202)	0.2(229)	0.0(135)	1.0(186)	1.5(259)	2.7(152)
IPTW oracle select and fuse	0.1(46)	0.4(36)	0.1(33)	4.7(202)	0.1(231)	0.1(135)	0.9(185)	0.9(266)	2.7(153)
LOAL	0.2(46)	0.5(37)	0.2(34)	6.0(190)	0.2(209)	0.5(123)	4.9(228)	5.6(284)	5.7(166)
Fused LOAL	0.1(46)	0.6(36)	0.2(34)	6.0(191)	0.2(213)	0.5(123)	4.9(228)	5.5(285)	5.6(166)
n=500									
G-comp main terms	0.1(55)	0(39)	0.1(33)	11.8(342)	0.8(292)	2.1(183)	45.5(1061)	32.1(798)	18.1(473)
IPTW full main terms	0.5(97)	0.4(82)	0.4(76)	6.1(478)	0.1(563)	0.3(303)	3.0(516)	4.2(673)	3.9(354)
IPTW oracle select	0.1(77)	0.2(67)	0.1(60)	5.3(395)	1.1(464)	0.3(280)	1.7(352)	3.6(503)	3.3(283)
IPTW oracle select and fuse	0.1(78)	0.2(66)	0.1(61)	5.3(392)	0.9(459)	0.3(279)	1.7(342)	3.2(495)	3.3(281)
LOAL	0.1(80)	0.4(63)	0.1(63)	7.2(375)	0.5(432)	0.3(255)	6.9(381)	7.0(537)	7.5(304)
Fused LOAL	0.0(81)	0.4(63)	-0.1(64)	7.3(372)	0.4(424)	0.3(253)	6.9(382)	6.8(542)	7.4(304)
n=1000									
G-comp main terms	0.1(81)	0.0(53)	0.0(48)	16.6(606)	1.5(411)	2.9(252)	63.9(2063)	43.4(1459)	25.5(871)
IPTW full main terms	0.3(148)	0.3(121)	0.2(117)	5.3(741)	0.0(1042)	0.2(517)	3.2(753)	2.6(1081)	4.7(546)
IPTW oracle select	0.1(115)	0.4(95)	0.1(87)	5.3(613)	0.9(778)	0.8(439)	1.1(524)	2.1(793)	3.7(420)
IPTW oracle select and fuse	0.0(116)	0.4(94)	0.1(88)	5.2(612)	0.9(784)	0.8(439)	1.1(524)	1.7(810)	3.6(424)
LOAL	0.1(117)	0.4(94)	0.1(91)	6.7(571)	0.2(726)	0.2(394)	8.1(590)	11.1(930)	8.9(479)
Fused LOAL	0.1(118)	0.4(94)	0.0(92)	6.8(568)	0.2(727)	0.2(394)	8.2(592)	10.8(936)	8.8(481)

Table 1: Scenario 1: $\sqrt{n}$ times the absolute value of bias (n times mean squared error) of methods estimating the parameters in the MSM of equation 1. IPTW oracle fits a glm of the target propensity score model with correctly selected variables and fused coefficients. The Fused LOAL uses the estimates of LOAL for the adaptive weights.

| n | Method | Selected covariates | | | | | | Fused non-zero coefficients | |
| | | A_0 model | | A_1 model | | | | | |
		C_0	I_0	C_0	I_0	C_1	I_1	C_0	I_0
Fusion results with oracle $\hat{\alpha}$ (independent of scenario)									
200	Fused LASSO	1.00	0.00	1.00	0.00	1.00	0.00	0.99	0.00
500	Fused LASSO	1.00	0.00	1.00	0.00	1.00	0.00	0.99	0.00
1000	Fused LASSO	1.00	0.00	1.00	0.00	1.00	0.00	0.99	0.00
a) Outcome generating model with main terms									
200	Fused LOAL	1.00	0.12	0.79	0.00	0.94	0.03	0.77	0.00
500	Fused LOAL	1.00	0.03	0.90	0.00	0.99	0.00	0.89	0.00
1000	Fused LOAL	1.00	0.01	0.95	0.00	1.00	0.00	0.94	0.00
	True βs	1.50	0.00	0.50	0.00	1.41	0.00		
	$\lim_{n\to\infty}\hat{\beta}$	1.50	0.00	0.50	0.00	1.65	0.00		
b) Outcome generating model with covariate interaction									
200	Fused LOAL	0.96	0.09	0.79	0.03	0.56	0.17	0.75	0.00
500	Fused LOAL	1.00	0.08	0.93	0.02	0.71	0.12	0.84	0.00
1000	Fused LOAL	1.00	0.06	0.98	0.01	0.84	0.14	0.83	0.00
	True βs	2.75	0.00	1.75	0.00	1.41	0.00		
	$\lim_{n\to\infty}\hat{\beta}$	2.70	0.00	1.73	0.00	1.64	-0.04		
c) Outcome generating model with effect modification									
200	Fused LOAL	1.00	0.16	0.24	0.03	1.00	0.21	0.22	0.00
500	Fused LOAL	1.00	0.13	0.32	0.02	1.00	0.17	0.28	0.00
1000	Fused LOAL	1.00	0.11	0.43	0.02	1.00	0.17	0.34	0.00
5000	Fused LOAL	1.00	0.06	0.78	0.01	1.00	0.16	0.60	0.00
	True βs	4.00	0.00	0.50	0.00	4.95	0.00		
	$\lim_{n\to\infty}\hat{\beta}$	5.40	0.04	0.61	0.05	7.82	0.09		

Table 2: Scenario 1: Proportion selection of each covariate into each treatment model and fusion of the coefficients of common terms across the two models, out of 1000 runs. C_0 and C_1 are true confounders and I_0 and I_1 are both instruments. The coefficients for C_0 are common across the two models under the target adjustment set. The true βs are the coefficients in the working models corresponding to equations (5) and (6). The $\lim_{n\to\infty}\hat{\beta}$ are the converging values of the estimates of β which depend on the specification of the q models. Fused LOAL was implemented in two ways: first, taking initial values of $\hat{\alpha}$ as estimated in a variable selection oracle model and secondly taking initial values from the LOAL (the former procedure being independent of the outcome).

schools in Montréal, Canada [O'Loughlin et al., 2015]. Self-report questionnaires were administered at three-month intervals, resulting in a total of 20 cycles from 1999 to 2005. An additional post-high school survey was conducted in 2007 or 2008. Data were collected from repeated assessments of a wide range of sociodemographic, substance use, psychosocial, lifestyle, and physical and mental health variables. We consider data from 1,231 students who were in grade seven in September 1999 and who were not previous regular (at least weekly use) alcohol users.

The baseline variables (cycle 1) included in our analysis were reported sex (female vs. male), mother's education (less than university vs. at least some university), single-parent home, French spoken at home, country of birth (outside Canada vs. Canada), self-esteem, impulsivity, and novelty-seeking. The time-varying covariates $\boldsymbol{L}$ considered were current depressive symptoms, participation in team sports, family-related stress, other type of stress, worry about weight, and ever smoked. The exposure A_t was the indicator of initiation of regular alcohol use at or before cycle t. Note that if $A_t = 1$ at a given time, then $A_k = 1$ at all times $k > t$ by our definition. We considered data from cycles 1 to 5, spanning calendar years 1999 to 2000, for the time-varying covariates and exposure. The outcome Y was depressive symptoms experienced within the past two weeks as measured using the Major Depressive Inventory (MDI) in 2007 or 2008 [Bech et al., 2015] when participants were age 20.4 years on average (i.e., approximately two years after cycle 20). The outcome is a continuous score ranging from 0 to 50, with higher scores indicating more severe symptoms. Since not all participants initially recruited took part in all cycles of the study, we denote loss to follow-up (i.e. censoring) by cycle t as $C_t = 1$, and $C_t = 0$ otherwise. The observed data structure is written as $O = (\boldsymbol{L}_1, A_1, \boldsymbol{L}_2, C_2, A_2, \cdots, A_5, \boldsymbol{L}_6, C_6, Y)$. Note that $\boldsymbol{L}_1$ contains the baseline covariates in addition to the time-varying covariates at the first time.

We denote an arbitrary exposure pattern as $\overline{\boldsymbol{a}} = (a_1, \cdots, a_5)$. Define $\mathcal{D}$ as the treatment regimen space, corresponding to the 6 possible treatment patterns: initiation at time 1, 2, 3, 4, or 5, or no initiation at any time point. For example, initiation at time 2 is represented as $(0, 1, 1, 1, 1)$. The parameters of interest were defined through the working MSM

$$\mathbb{E}[Y^{\overline{\boldsymbol{a}}}|\text{Sex}] = \mu_0 + \mu_1 \text{Sex} + \mu_2 cum(\overline{\boldsymbol{a}}) + \mu_3\{\text{Sex} \times cum(\overline{\boldsymbol{a}})\}, \tag{8}$$

where $cum(\overline{a})$ gives the number of exposed time points in $\overline{a}$.

We extended out methods to incorporate the simultaneous presence of time-dependent treatment and censoring. We considered four implementations of IPTW and LTMLE:

- IPTW full: Fit stratified models for the probability of being exposed (and censored, respectively) at each given time according to all previous covariates' main terms among participants who were previously unexposed and uncensored.

- IPTW LOAL and IPTW fused LOAL: Included selected variables in the treatment and censoring models using the LOAL and fused LOAL procedures, respectively.

- LTMLE full: Included all covariates in the stratified treatment and censoring models.

- LTMLE LOAL and LTMLE fused LOAL: Included only the covariates selected using the LOAL and fused LOAL procedures, respectively, in the treatment and censoring models.

For all implementations, the outcome models included the main terms of the baseline and time-varying covariates, current and lagged exposure terms, and the first-order interactions of sex and exposures for uncensored participants.

For LOAL and fused LOAL, we performed variable selection and fusion for the treatment model and censoring model separately, but the penalization parameters λ_n^a (for treatment) and λ_n^c (for

censoring) were selected jointly by minimizing the sum of two longitudinal balancing metrics over covariates and times with respect to treatment and censoring at each time point. For fused LOAL, the penalty graph connected common baseline variables across time points, as well as common time-varying variables with the same lag across time points (e.g., corresponding $\boldsymbol{L}_{t-1}$ variables are connected together when modeling A_t, and when modeling C_t).

We fixed the tuning parameter γ at 2.5 and used 20 candidate values for the tuning parameters λ^a with the range (e^{-4}, e^8) and λ^c with the range (e^{-5}, e^{10}), with values increasing evenly on a log scale, and which, at the extremes, included both null and complete variable selection and fusing.

The full treatment model included 135 parameters (including five intercepts) and was reduced to 37 parameters by LOAL, where the variables sex, and current depressive symptoms were selected in each time period. The fusion step further reduced the number of parameters to 23, a reduction of 83% in the number of parameters as compared to the full model (see Table 3). The full censoring model included 180 parameters (including five intercepts and 15 coefficients of past treatments), of which 112 remained after LOAL and 55 after fused LOAL, representing a reduction of 69% of the number of parameters. The variables selected in the censoring model included sex, country of birth, current depressive symptoms, ever smoked, family-related stress, other stress, participation in team sports and worry about weight (see Table 4).

Variable \ Time	1	2	3	4	5
Intercept	-4.067	-3.475	-3.898	-3.609	-1.469
Sex	-0.313	-0.313	-0.313	-0.313	-0.313
CountryBirth	-0.838	-0.838		-0.838	-0.838
MotherEducation					0.491
$\text{CurDep}_{t=1}$	1.493				
$\text{CurDep}_{t=2}$	NA	1.493			
$\text{CurDep}_{t=3}$	NA	NA	1.493	0.127	
$\text{CurDep}_{t=4}$	NA	NA	NA	1.493	0.127
$\text{CurDep}_{t=5}$	NA	NA	NA	NA	1.493
$\text{EverSmoke}_{t=3}$	NA	NA	1.349	1.455	0.916
$\text{FamStress}_{t=2}$	NA				-0.208
$\text{FamStress}_{t=4}$	NA	NA	NA		0.072
$\text{OtherStress}_{t=5}$	NA	NA	NA	NA	0.081
$\text{TeamSport}_{t=3}$	NA	NA		-0.020	0.147
$\text{TeamSport}_{t=5}$	NA	NA	NA	NA	0.117
$\text{WorWeight}_{t=1}$		0.277		0.410	0.074
$\text{WorWeight}_{t=3}$	NA	NA	-0.119		
$\text{WorWeight}_{t=4}$	NA	NA	NA	-0.119	0.277

Table 3: Selected and fused parameters in the treatment model. 'CurDep' represents current depressive symptoms; 'FamStress' represents family stress; 'TeamSport' represents participation in team sports; 'WorWeight' represents worry about weight; 'NA' represents that the time varying variable is not applicable at the given time. A blank space means that the variable was not selected by the LOAL in the first step. The values are color coded such that common colors indicate fused parameters.

Variable \ Time	2	3	4	5	6
Intercept	-4.980	-4.106	-2.445	-1.064	-1.443
A1	2.458			0.423	-0.810
A2	NA		-1.395	0.037	-1.429
A3	NA	NA	0.840	-0.594	3.187
A4	NA	NA	NA	0.709	-1.946
A5	NA	NA	NA	NA	-0.097
Sex	-0.336	-0.336	-0.336	-0.336	-0.336
SelfEsteem	-0.133	-0.133	-0.133	-0.133	
CountryBirth	0.760	0.760	0.760	0.760	0.760
MotherEducation			-0.255	-0.255	
$\text{CurDep}_{t=1}$			-0.823	-0.360	-0.263
$\text{CurDep}_{t=2}$	-0.242	-0.167	0.358		-0.360
$\text{CurDep}_{t=3}$	NA	-0.242	-0.167	0.358	
$\text{CurDep}_{t=4}$	NA	NA	-0.242	-0.167	0.358
$\text{CurDep}_{t=5}$	NA	NA	NA	-0.242	-0.167
$\text{CurDep}_{t=6}$	NA	NA	NA	NA	-0.242
$\text{EverSmoke}_{t=1}$		0.235	0.264	-0.001	0.568
$\text{EverSmoke}_{t=2}$		-0.192	0.235	0.264	-0.001
$\text{EverSmoke}_{t=3}$	NA	0.506	-0.192	0.235	
$\text{EverSmoke}_{t=4}$	NA	NA	0.506	-0.192	
$\text{EverSmoke}_{t=5}$	NA	NA	NA	0.506	
$\text{EverSmoke}_{t=6}$	NA	NA	NA	NA	0.506
$\text{FamStress}_{t=2}$	0.022	0.044	-0.039	0.329	-0.246
$\text{FamStress}_{t=3}$	NA		0.044	-0.039	
$\text{FamStress}_{t=4}$	NA	NA		0.044	-0.039
$\text{FamStress}_{t=5}$	NA	NA	NA	0.022	0.044
$\text{FamStress}_{t=6}$	NA	NA	NA	NA	0.022
$\text{OtherStress}_{t=1}$	0.070	-0.219	0.353	0.150	-0.054
$\text{OtherStress}_{t=3}$	NA			-0.219	
$\text{OtherStress}_{t=4}$	NA	NA	-0.025	0.070	
$\text{OtherStress}_{t=5}$	NA	NA	NA	-0.025	0.070
$\text{TeamSport}_{t=2}$	-0.105	-0.185	0.128	0.004	0.191
$\text{TeamSport}_{t=3}$	NA		-0.185	0.128	0.004
$\text{TeamSport}_{t=4}$	NA	NA	-0.105	-0.185	
$\text{TeamSport}_{t=5}$	NA	NA	NA	-0.105	-0.185
$\text{TeamSport}_{t=6}$	NA	NA	NA	NA	-0.105
$\text{WorWeight}_{t=1}$	-0.018	0.053	0.044	-0.388	0.012
$\text{WorWeight}_{t=2}$	0.145				
$\text{WorWeight}_{t=3}$	NA	0.145			
$\text{WorWeight}_{t=4}$	NA	NA	0.145	-0.018	0.053
$\text{WorWeight}_{t=5}$	NA	NA	NA	0.145	-0.018
$\text{WorWeight}_{t=6}$	NA	NA	NA	NA	0.145

Table 4: Selected and fused parameters in the censoring model. 'CurDep' represents current depressive symptoms; 'FamStress' represents family stress; 'TeamSport' represents participation in team sports; 'WorWeight' represents worry about weight; 'na' represents that the time-dependent variable is not applicable at the given time. A blank space means that the variable was not selected by the LOAL in the first step. The values are color coded according to fused clique.

Estimated coefficients and standard errors are presented in Section 6. The standard errors of the three LTMLE estimates were obtained through the influence functions without accounting for variable selection and are thus not valid post-inference standard errors. For IPTW, we applied the robust sandwich variance estimator. All methods consistently indicated that females had more severe depressive symptoms compared to males. Point estimates from IPTW full, LTMLE full, LTMLE LOAL, and LTMLE fused LOAL suggested that early alcohol initiation was associated with more depressive symptoms during young adulthood among male participants. All IPTW implementations suggested that alcohol initiation in female participants was associated with less severe depressive symptoms; however, the LTMLE implementations concluded null or harmful impacts of earlier drinking initiation for both sexes. Notably, the incorporation of propensity scores limited to covariates selected by LOAL led to greatly reduced variance estimates in both the IPTW and LTMLE analyses. In addition, LOAL plus fusion more than halved the estimated variance of the LTMLE estimator compared to LTMLE with only LOAL.

Method\Coefficient	Intercept	Female sex	$\text{cum}(\overline{\boldsymbol{a}})$	Female sex$\times\text{cum}(\overline{\boldsymbol{a}})$
IPTW full	7.397(1.119)	5.708(2.177)	0.583(0.646)	-1.833(1.564)
IPTW LOAL	8.203(0.975)	5.774(1.898)	-0.224(0.438)	-0.352(0.844)
IPTW fused LOAL	8.553(0.943)	4.223(1.452)	-0.047(0.636)	-0.455(1.038)
LTMLE full	7.361(0.337)	3.569(0.573)	0.072(0.253)	0.008(0.504)
LTMLE LOAL	7.570(0.104)	3.479(0.182)	0.005(0.085)	0.028(0.154)
LTMLE fused LOAL	7.712(0.028)	3.500(0.045)	0.002(0.042)	-0.008(0.076)

Table 5: Estimates of the MSM parameters for the NDIT study application. Estimated standard errors are presented in brackets. IPTW [LTMLE] full represents IPTW [LTMLE] with pooled treatment models including all covariate main terms and pooled censoring models including all covariate main terms and treatment terms; IPTW [LTMLE] LOAL represents IPTW [LTMLE] with pooled treatment and censoring models after covariate selection by LOAL; IPTW [LTMLE] fused LOAL represents IPTW [LTMLE] with pooled treatment and censoring models with both selection by LOAL and coefficient fusion.

Complete details of the application and results are given in Web Appendix D.

7 Discussion

In this paper, we extended the Outcome Adaptive LASSO propensity score variable selection approach of Shortreed and Ertefaie [2017] to the setting with time-varying treatment over discrete time points. We first estimate regularized coefficients of the time-saturated propensity score models. We then fuse the resulting nonzero coefficients using a generalized adaptive fused LASSO. Allowing for sparse model identification can avoid forcing a Markov-type assumption where we assume a priori that treatment can only depend on the most recent values of the time-updated covariates and baseline covariates. Oracle properties of the generalized adaptive fused LASSO [Viallon et al., 2016] guarantee oracle performance of this estimator. In our setting, this means that the fused LASSO will fuse the coefficients correctly according to the marginal pooled treatment model as sample size increases.

Our simulation studies show that implementation of our method can improve estimation by IPTW and LTMLE compared to the same estimators without variable selection. Our approach also

16

exhibited better performance than G-computation and C-LTMLE under incorrect outcome model specification. The success of the selection relied on the specification of the outcome model to the extent that the estimates of the βs in the working model converged correctly to either zero or non-zero; however, the success of variable selection also depended on the variance of the estimators of these βs. The fusion was highly successful whenever the correct covariates were selected in the LOAL step.

The application demonstrated the usage of our method in a realistically complex longitudinal study in epidemiology where the interest was in estimating the effect of the initiation time of regular alcohol consumption on depression symptoms in young adulthood. We extended our method to select adjustment variables in both the treatment and censoring models, and used the balance criterion to jointly select the tuning parameter for the two models. The reduction of covariates and fusion of coefficients in the treatment and censoring models both led to apparent major gains in efficiency.

An important limitation of all frequentist covariate selection methods for the propensity score that exclude instruments is that inference is not formally available [Leeb and Pötscher, 2005, Tang et al., 2022]. This means that in practice, it may be best to avoid covariate selection in combination with IPTW and LTMLE if possible. Another limitation of our method, as currently proposed, is that it cannot handle nonlinearities or interactions between covariates in the propensity score models. Potential extensions of our method may incorporate nonparametric approaches used in the single time-point setting such as causal ball screening [Tang et al., 2022] for high-dimensional covariates and outcome highly adaptive lasso [Ju et al., 2020], the latter of which has valid closed-form expressions for confidence intervals. To conclude, we consider this work as a step in the development of nonparametric shrinkage estimators that trade-off bias and variance in the MSM parameter estimation for longitudinal treatments while allowing for valid inference.

Acknowledgements

Supporting Information

Web Appendencies, tables, and figures referenced in Sections 3.4, 4.2, 5.1, 5.2, and 6 are available with this paper at the Biometrics website on Oxford Academic. The R codes used for the simulations are available on github [to be uploaded, currently attached to submission].

References

Please find the references at the end of this document.

Supplementary Materials for "Adaptive sparsening and smoothing of the treatment model for longitudinal causal inference using outcome-adaptive LASSO and marginal fused LASSO"

by Mireille E Schnitzer, Denis Talbot, Yan Liu, David Berger, Guanbo Wang, Jennifer O'Loughlin, Marie-Pierre Sylvestre, and Ashkan Ertefaie

A Balance metric for variable selection

Our tuning parameter selection for the LOAL evaluates the balance of every covariate or function of covariates at each time point between units with different current treatment values. It is similar to a summary of the longitudinal balancing metric in Jackson [2016] (Diagnostic 3). However, unlike Jackson [2016], we do not subset on treatment history as this will result in vanishing data support for greater numbers of time points; our definition of the weights in the main manuscript, which does not stabilize conditional on past treatments, produces independence between past treatment and current covariates. Adenyo et al. [2024a] Each covariate $L_{t,k}$ considered at each time point $\tau \geq t$ will be weighted by the corresponding structural model coefficient ($\hat{\beta}_{\tau,t,k}$) divided by its standard error ($\sigma_{\hat{\beta}_{\tau,t,k}}$). We let $\hat{\boldsymbol{\alpha}}^{refit}(\lambda_n)$ represent the estimated values of the $\boldsymbol{\alpha}$ parameters using a logistic regression on the covariates selected by LOAL under tuning parameter value λ_n, where the value is defined as zero if the corresponding coefficient was not selected. Then, define the LOAL-weight for subject i at time τ as $\hat{w}_{\tau,i}^{\lambda_n}(\bar{a}_\tau) = w_\tau\{\bar{a}_\tau, \bar{\boldsymbol{l}}_{\tau,i}; \hat{\boldsymbol{\alpha}}^{refit}(\lambda_n)\}$ corresponding to the weights defined in the main manuscript. For the simple example, we use the metric

$$
\begin{aligned}
\sum_{k=1}^{p_0} & \frac{|\hat{\beta}_{0,0,k}|}{\sigma_{\hat{\beta}_{0,0,k}}} \left| \frac{\sum_{i=1}^n a_{0,i} l_{0,k,i} \hat{w}_{0,i}^{\lambda_n}(1)}{\sum_{i=1}^n a_{0,i} \hat{w}_{0,i}^{\lambda_n}(1)} - \frac{\sum_{i=1}^n (1-a_{0,i}) l_{0,k,i} \hat{w}_{0,i}^{\lambda_n}(0)}{\sum_{i=1}^n (1-a_{0,i}) \hat{w}_{0,i}^{\lambda_n}(0)} \right| \\
& + \sum_{k=1}^{p_0} \frac{|\hat{\beta}_{1,0,k}|}{\sigma_{\hat{\beta}_{1,0,k}}} \left| \frac{\sum_{i=1}^n a_{1,i} l_{0,k,i} \hat{w}_{1,i}^{\lambda_n}(a_{0,i},1)}{\sum_{i=1}^n a_{1,i} \hat{w}_{1,i}^{\lambda_n}(a_{0,i},1)} - \frac{\sum_{i=1}^n (1-a_{1,i}) l_{0,k,i} \hat{w}_{1,i}^{\lambda_n}(a_{0,i},0)}{\sum_{i=1}^n (1-a_{1,i}) \hat{w}_{1,i}^{\lambda_n}(a_{0,i},0)} \right| \\
& + \sum_{k=1}^{p_1} \frac{|\hat{\beta}_{1,1,k}|}{\sigma_{\hat{\beta}_{1,1,k}}} \left| \frac{\sum_{i=1}^n a_{1,i} l_{1,k,i} \hat{w}_{1,i}^{\lambda_n}(a_{0,i},1)}{\sum_{i=1}^n a_{1,i} \hat{w}_{1,i}^{\lambda_n}(a_{0,i},1)} - \frac{\sum_{i=1}^n (1-a_{1,i}) l_{1,k,i} \hat{w}_{1,i}^{\lambda_n}(a_{0,i},0)}{\sum_{i=1}^n (1-a_{1,i}) \hat{w}_{1,i}^{\lambda_n}(a_{0,i},0)} \right|.
\end{aligned}
$$

B Asymptotics of the Generalized Adaptive Fused LASSO

Here we state a theorem on the convergence of the estimates of the generalized adaptive fused LASSO that does not include a sparsity-inducing penalty and is performed after variable selection. The result is directly connected to the main result in Viallon et al Viallon et al. [2016]. We present this statement with the purpose of giving details about the theoretical properties of the second stage of our procedure.

Recall that the parameter $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{0,-1}, \boldsymbol{\alpha}_{0,0}, \boldsymbol{\alpha}_{1,-1}, \boldsymbol{\alpha}_{1,0}, \boldsymbol{\alpha}_{1,1}, \boldsymbol{\alpha}_{1,2})$ is defined according to the model (4) in the main manuscript. Define $\boldsymbol{\alpha}^\dagger = (\alpha_{0,-1}, \boldsymbol{\alpha}_{0,0}^\dagger, \alpha_{1,-1}, \boldsymbol{\alpha}_{1,0}^\dagger, \boldsymbol{\alpha}_{1,1}^\dagger, \alpha_{1,2})$, a parameter vector potentially containing fixed zeros among the elements of $\boldsymbol{\alpha}_{0,0}^\dagger$, $\boldsymbol{\alpha}_{1,0}^\dagger$, and $\boldsymbol{\alpha}_{1,1}^\dagger$. The elements that are not fixed zeros are free parameters. Suppose that the marginalized distribution of the treatments $A_t; t = 0, 1$ corresponds to Bernoulli distributions with probability of success $m_\tau(\overline{L}_1, A_0; \boldsymbol{\alpha}_0^\dagger)$ where $\boldsymbol{\alpha}^{\dagger,*}$ are defined as the true parameter values under maximum likelihood. More specifically, the distribution of A_0 conditional on the elements of L_0 corresponding to non-fixed-zero components

of $\boldsymbol{\alpha}_{0,1}^{\dagger}$ is Bernoulli with probability of success $m_0(\overline{L}_1, A_0; \boldsymbol{\alpha}^{\dagger,*})$; the distribution of A_1 conditional on A_0 and the elements of (L_0, L_1) corresponding to non-fixed-zero components of $(\boldsymbol{\alpha}_{1,0}^{\dagger}, \boldsymbol{\alpha}_{1,1}^{\dagger})$ is Bernoulli with probability of success $m_1(\overline{L}_1, A_0; \boldsymbol{\alpha}^{\dagger,*})$.

Let $\mathcal{J}_{0,1}^{\dagger}$ and $\mathcal{J}_{1,1}^{\dagger}$ denote the indices of the covariates corresponding to the non-zero elements of $\boldsymbol{\alpha}_{0,0}^{\dagger}$ and $\boldsymbol{\alpha}_{1,0}^{\dagger}$, respectively. We define a graph $\mathcal{G} = (V, E)$ with vertices $V = \{(0, 0, k_0), (1, 0, k_1); k_0 \in \mathcal{J}_{0,0}^{\dagger}$ and $k_1 \in \mathcal{J}_{1,0}^{\dagger}\}$ and edges E that connect all (pairs of) corresponding indices for $k \in \mathcal{J}_{0,0}^{\dagger} \cap \mathcal{J}_{1,0}^{\dagger}$. This is the graph that will be used to run the adaptive fused LASSO.

Define the estimator $\hat{\boldsymbol{\alpha}}^{\dagger}$ as the minimizer of

$$\sum_{\tau=0}^{1} \sum_{i=1}^{n} \left[a_{\tau,i} \log\{m_\tau(\bar{l}_{1,i}, a_{0,i}; \boldsymbol{\alpha}^{\dagger})\} + (1 - a_{\tau,i}) \log\{1 - m_\tau(\bar{l}_{1,i}, a_{0,i}; \boldsymbol{\alpha}^{\dagger})\} \right]$$
$$+ \lambda_{1,n} \sum_{k \in \mathcal{J}_{0,0}^{\dagger} \cap \mathcal{J}_{1,0}^{\dagger},} \frac{|\alpha_{1,0,k}^{\dagger} - \alpha_{0,0,k}^{\dagger}|}{|\tilde{\alpha}_{1,0,k}^{\dagger} - \tilde{\alpha}_{0,0,k}^{\dagger}|^{\gamma_1}} \tag{9}$$

where $\tilde{\alpha}_{1,0,k}^{\dagger}$ and $\tilde{\alpha}_{0,0,k}^{\dagger}$ are $\sqrt{n}$-consistent estimates of $\alpha_{1,0,k}^{\dagger,*}$ and $\alpha_{0,0,k}^{\dagger,*}$, respectively.

Now following Viallon et al. [2016] with adaptations to our setting, we define $\mathcal{J}^* = \{k \in \mathcal{J}_{0,1}^{\dagger} \cap \mathcal{J}_{1,1}^{\dagger} : \alpha_{0,1,k}^{\dagger,*} = \alpha_{1,1,k}^{\dagger,*}\}$, i.e. the set of indices where fusing should occur. Furthermore, let $\mathcal{B} = \{(0, 1, k), (1, 1, k) \in E : j \in \mathcal{J}^*\} \subseteq E$, which is a set of connected indices where the true values of the parameters are equal. Define the graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{B})$ as the one containing the complete set of vertices of $\mathcal{G}$ but with edges only between the connected vertices of E where the corresponding parameters values are equal. Define s_0 as the number of connected components of the graph $\mathcal{G}_{\mathcal{B}}$. Now define $\alpha_{\mathcal{B}}^* = (\alpha_{0,0}^*, \alpha_{0,1}^{\dagger,*}, \alpha_{1,0}^*, \alpha_{1,1,\mathcal{J}_{1,1} \backslash \mathcal{J}^*}^*, \alpha_{1,2}^{\dagger,*}, \alpha_{1,3}^*)^T$, which are the same as $\alpha^{\dagger,*}$ after removal of the redundant terms in $\alpha_{1,1}$ that are equal to their connected terms in $\alpha_{0,1}$; and let $\hat{\alpha}_{\mathcal{B}}$ be its estimate by the adaptive fused LASSO in (9). Finally, define $\mathcal{B}_n = \{(0, 1, k), (1, 1, k) \in E : \hat{\alpha}_{0,1,k}^{\dagger} = \hat{\alpha}_{1,1,k}^{\dagger}\}$, the edges that fused in the procedure.

Theorem 1 *If $\lambda_{1,n}/\sqrt{n} \to 0$ and $\lambda_{1,n} n^{(\gamma_1 - 1)/2} \to \infty$, then under mild assumptions, the minimizer of (9) satisfies $P(\mathcal{B}_n = \mathcal{B}) \to 1$ as $n \to \infty$ and $\sqrt{n}(\hat{\alpha}_{\mathcal{B}} - \alpha_{\mathcal{B}}^*)$ converges in distribution to a Gaussian distribution of dimension $s_0 + 3$ with mean zero.*

The mild assumptions are given explicitly in Viallon et al Viallon et al. [2013] as AL1 and AL2. The proof of our theorem follows exactly the steps of their proof of Theorem 2 excluding the sparsity element.

C   Simulation study details and extended results

C.1   Scenario 1 data generating mechanisms

The data in Scenario 1 were generated according to the left-hand DAG in Figure 2 and more specifically from the mechanisms presented in Table 6.

Table 6: Simulation Scenario 1 data generating mechanism

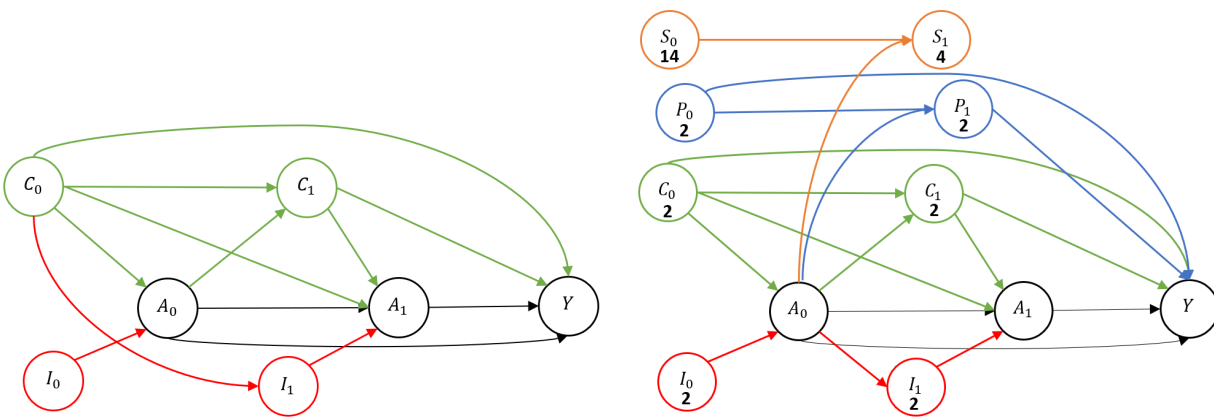| Variable	Generating Mechanism
C_0	$\sim N(\text{mean} = 0, \text{sd} = 1)$
I_0	$\sim N(\text{mean} = 0, \text{sd} = 1)$
A_0	$\sim \text{Bernoulli}(\text{logit}(p) = 1.515C_0 + I_0)$
C_1	$\sim N(\text{mean} = C_0 + A_0, \text{sd} = 1)$
I_1	$\sim N(\text{mean} = C_0, \text{sd} = 1)$
A_1	$\sim \text{Bernoulli}(\text{logit}(p) = -0.5 + 0.5C_0 + 0.25C_1 + 0.5A_0 + I_1)$
Y for scenario (a)	$\sim N(\text{mean} = -1.5 + 0.5C_0 + 0.5A_0 + C_1 + A_1, \text{sd} = 0.5)$
Y for scenario (b)	$\sim N(\text{mean} = -1.5 + 0.5C_0 + 0.5A_0 + C_1 + A_1 + 2.5C_0C_1, \text{sd} = 0.5)$
Y for scenario (c)	$\sim N(\text{mean} = -1.5 + 0.5C_0 + 0.5A_0 + C_1 + A_1 + 2.5A_0C_1^2, \text{sd} = 0.5)$

Figure 2: DAGs representing the data generation in Scenarios 1 (left) and 2 (right). The target variable selection retained all covariates labeled C and P (the latter only for Scenario 2 in both models).

C.2 Scenario 2: higher dimensional covariates with two time-points

In this scenario, we generated 20 independent covariates at time 0 and 10 covariates at time 1, two treatments and a continuous outcome, according to the right-hand DAG in Figure 2 and more specifically the complete data generating mechanism in Table 7. At time 0 there were two confounders, jointly denoted C_0; two pure causes of the outcome, P_0; two instruments, I_0, and 14 spurious covariates S_0. At time 1 there were two confounders, C_1; two pure causes of the outcome, P_1, two instruments, I_1, and four spurious covariates S_1. All of the time 1 covariates were affected by the corresponding covariate at time 0 and also by the previous treatment A_0. The data were generated in such a way that the coefficients of both confounders in C_0 at the two time-points were equal in the marginal pooled treatment model that excluded instruments and spurious covariates. The coefficients of the two variables P_0 were equal to zero in this same model. The outcome was Gaussian with mean linear in the main terms of C, P, A_0, and A_1. The model to estimate q_1 conditioned on all main terms (and thus contained the truth) for all methods; the models for q_0 were linear in the main terms.

The $\sqrt{n}$-bias and n-MSE are given in the first three data columns of Table 9. G-computation was unbiased with the lowest MSE as it was approximately correctly specified. The oracle IPTWs

Table 7: Simulation Scenario 2 data generating mechanism

Variable	Generating Mechanism
$C_{0,j}$ for $j = (1,2)$	$\sim N(\mathsf{mean} = 0, \mathsf{sd} = 1)$
$P_{0,j}$ for $j = (1,2)$	$\sim N(\mathsf{mean} = 0, \mathsf{sd} = 1)$
$I_{0,j}$ for $j = (1,2)$	$\sim N(\mathsf{mean} = 0, \mathsf{sd} = 1)$
$S_{0,j}$ for $j = (1,...,14)$	$\sim N(\mathsf{mean} = 0, \mathsf{sd} = 1)$
A_0	$\sim \mathsf{Bernoulli}(\mathsf{logit}(p) = C_{0,1} + C_{0,2} + I_{0,1} + I_{0,2})$
$C_{1,1}$	$\sim N(\mathsf{mean} = 0.5C_0 + 0.5A_0, \mathsf{sd} = 1)$
$C_{1,2}$	$\sim N(\mathsf{mean} = 0.2C_0 - A_0, \mathsf{sd} = 1)$
$P_{1,1}$	$\sim N(\mathsf{mean} = 0.5C_0 + 0.5A_0, \mathsf{sd} = 1)$
$P_{1,2}$	$\sim N(\mathsf{mean} = 0.2C_0 - A_0, \mathsf{sd} = 1)$
$I_{1,1}$	$\sim N(\mathsf{mean} = -0.5A_0, \mathsf{sd} = 1)$
$I_{1,2}$	$\sim N(\mathsf{mean} = A_0, \mathsf{sd} = 1)$
$S_{1,j}$ for $j = (1,...,4)$	$\sim N(\mathsf{mean} = 0.5C_0 + 0.2A_0, \mathsf{sd} = 1)$
A_1	$\sim \mathsf{Bernoulli}(\mathsf{logit}(p) = 1.026C_{0,1} + 0.987C_{0,2} + 0.5A_0 + C_{1,1} + C_{1,2} + I_{1,1} + I_{1,2})$
Y	$\sim N(\mathsf{mean} = 1 + 0.6C_{0,1} + 0.6C_{0,2} + 0.6P_{0,1} + 0.6P_{0,2} + 0.6C_{1,1} + 0.6C_{1,2}$ $+ 0.6P_{1,1} + 0.6P_{1,2} + 0.5A_0 + A_1, \mathsf{sd} = 1)$

produced lower bias and MSE than the full IPTW. While the LOAL and fused LOAL had higher bias than their oracle counterparts, they had lower MSE. Figure 3 gives the proportion selection for each covariate at each time point and proportion fused (and non-zero) for corresponding baseline covariates between the two time points. While the confounders C were selected nearly 100% of the time at all sample sizes, the selection of P varied between roughly 75-100%, and appeared to be slowly converging. The selection of instruments I varied between 10-20% and appeared to be slowly converging to zero. Spurious covariates S were selected less often (below 10%) and converged close to zero by $n = 1000$. The fusion of both C_0 variables quickly converged to almost 100% by $n = 1000$. The terms P_0 often fused when they were both selected and non-zero, about 60-75% of the time.

C.3   Scenario 3: five time-points with only baseline covariates

Finally, in order to demonstrate the potential usefulness of fusing, we developed a scenario with five treatments over time. Using the same approach as in scenario 2, we simulated 20 baseline covariates (two confounders C_0, two pure predictors of the outcome P_0, two instruments I_0, and 14 spurious covariates S_0). To facilitate the construction of this scenario, we did not generate covariates past the baseline. The data generation was done in such a way that, in the marginal pooled treatment model, all 5 of the coefficients for each confounder were equal, and all 10 of the outcome predictor coefficients were equal to zero. Thus, variable selection should reduce the number of parameters (including 5 intercepts and 10 coefficients of treatment) from 115 to 35, and fusion should further reduce the number of parameters to 19. See Table 8 for the complete data generating mechanism.

The $\sqrt{n}$-bias and n-MSE results are given in the last three columns of Table 9. G-computation was approximately correctly specified and unbiased. The full IPTW was more biased. IPTW with oracle variable selection decreased the bias and MSE relative to the full IPTW. IPTW with oracle variable selection and fusing produced smaller MSE compared to IPTW with oracle variable selection. Fused LOAL produced reductions in bias for μ_1 and MSE reductions for μ_0 and μ_1 compared to LOAL.
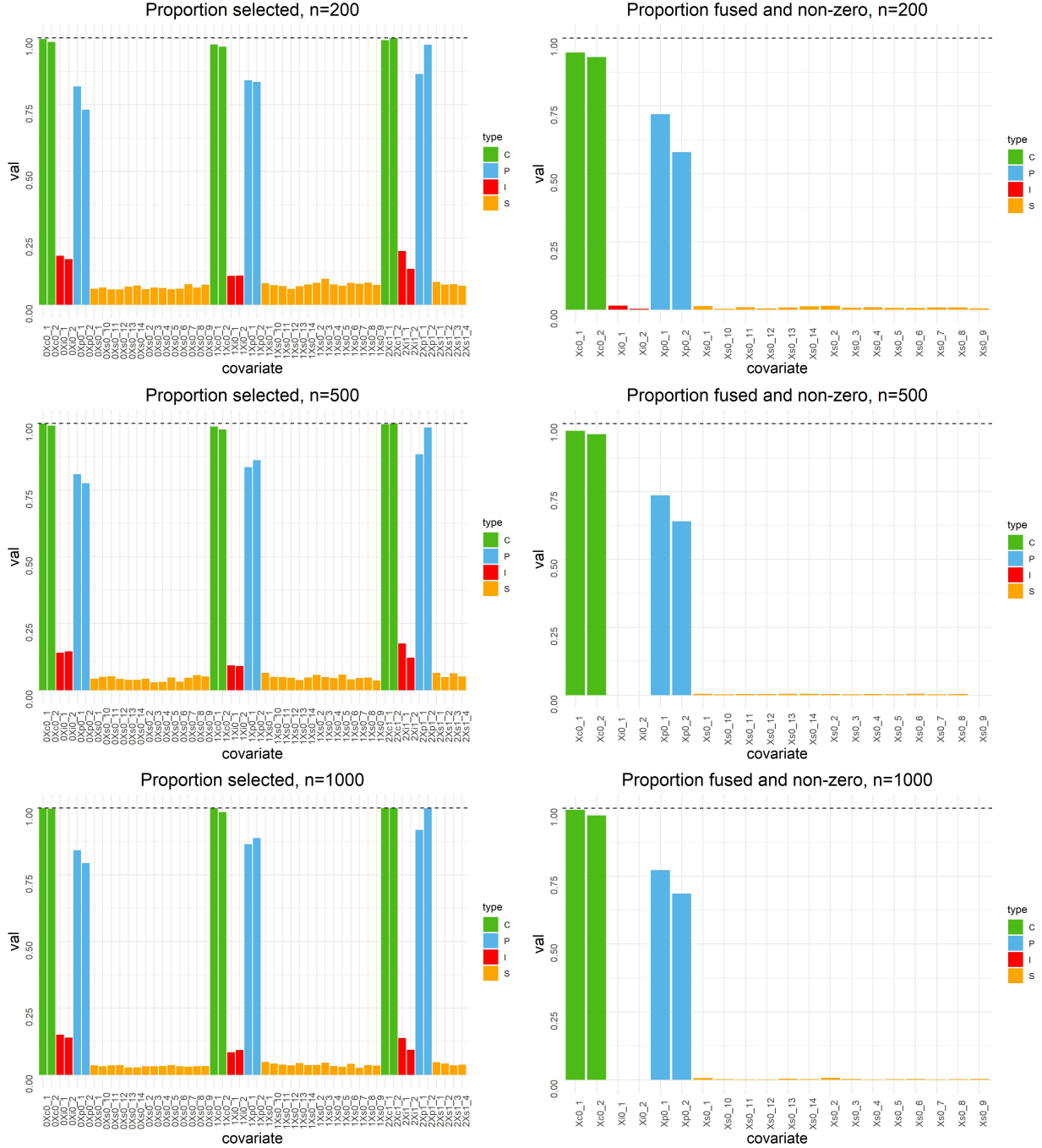
Figure 3: Proportion variable selection (left) and fusion (right) for $n = 200$ (top), $n = 500$ (middle), and $n = 1000$ (bottom) in Scenario 2.

The selection and fusion results are given in Figure 4. The confounder variable selection was again close to ideal. The selection of pure causes of the outcome increased to above 90% as n increased. Instrument and spurious covariate selection was again low. We also verified that the fusion of non-zero coefficients of the confounders matched the selection of confounders and was near perfect for $n = 1000$.

Table 8: Simulation Scenario 3 data generating mechanism

Variable	Generating Mechanism
$C_{0,j}$ for $j=(1,2)$ $P_{0,j}$ for $j=(1,2)$ $I_{0,j}$ for $j=(1,2)$ $S_{0,j}$ for $j=(1,...,14)$	$\sim$ Multivariate normal distribution with mean 0 and the covariance matrix $$\Sigma_{(20\times20)} = \begin{bmatrix} 0.64 & 0.192 & 0.192 & \cdots & 0.192 \\ 0.192 & 0.64 & 0.192 & \cdots & 0.192 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.192 & 0.192 & 0.192 & \cdots & 0.64 \end{bmatrix}$$
A_0	$\sim$ Bernoulli(logit$(p) = 0.5C_{0,1} + C_{0,2} - 0.5I_{0,1} - 0.5I_{0,2}$)
A_1	$\sim$ Bernoulli(logit$(p) = 0.542C_{0,1} + 1.075C_{0,2} - 0.545I_{0,1} - 0.545I_{0,2} - 0.5A_0$)
A_2	$\sim$ Bernoulli(logit$(p) = 0.568C_{0,1} + 1.142C_{0,2} - 0.565I_{0,1} - 0.569I_{0,2} - 0.5A_1$)
A_3	$\sim$ Bernoulli(logit$(p) = 0.615C_{0,1} + 1.23C_{0,2} - 0.61I_{0,1} - 0.61I_{0,2} - 0.5A_2$)
A_4	$\sim$ Bernoulli(logit$(p) = 0.66C_{0,1} + 1.322C_{0,2} - 0.655I_{0,1} - 0.655I_{0,2} - 0.5A_3$)
Y	$\sim N$(mean $= 0.6C_{0,1} + 0.6C_{0,2} + 0.6P_{0,1} + 0.6P_{0,2} + 0.5A_0 + 0.5A_1 + 0.5A_2$ $+0.5A_3 + 0.5A_4$, sd $= 1$)

C.4 Comparison with C-LTMLE

We used the simulation Scenarios 1 and 2 to contrast the estimation of $\mathbb{E}(Y^{1,1})$ by LTMLE using variable selection with LOAL to an implementation of C-LTMLE, which was implemented for two time-points and the interventional mean in previous work. Schnitzer et al. [2020] This C-LTMLE procedure uses initial estimates of $q_t; t = 1, 0$ to select covariates into the two propensity score models in a greedy fashion, creating a sequence of nested models with corresponding TMLE-updated estimates of $q_t; t = 0, 1$ with uniformly decreasing risk. Cross-validation is then used to select the point at which to stop, and this number of selections is used to determine the chosen propensity score models. (See Gruber and van der Laan [2010] and Schnitzer et al. [2020] for more details.) Using the same linear main-terms models to estimate q_t throughout, we fit 1) LTMLE using propensity scores conditional on the full set of covariates, 2) C-LTMLE, and 3) LTMLE using propensity scores with the covariates selected by LOAL.

The results in Table 10 show that both C-LTMLE and LTMLE with LOAL improved the bias and MSE over LTMLE implemented with all covariates. In Scenarios 1(a) and 2 when the outcome models were nearly correctly specified, C-LTMLE generally outperformed LTMLE with LOAL. However, with incorrect model specifications (Scenarios 1(b) and 1(c)), LTMLE LOAL had less bias and MSE. Compared to LTMLE with LOAL, C-LTMLE was very slow; for example, in Scenario 2 with $n = 1000$, C-LTMLE had a 90 second runtime while LTMLE with LOAL took only 3 seconds. However, the C-LTMLE procedure may be made more scalable with a preliminary ordering of the variables rather than the greedy procedure we employed. Ju et al. [2019]

Table 9: Scenarios 2 and 3 $\sqrt{n}$ times the absolute value of bias (n times mean squared error) of methods estimating the parameters in the marginal structural model of equation (see Equation 1 in Section 2.1 of the manuscript). The Fused LOAL uses the estimates of LOAL for the adaptive weights.

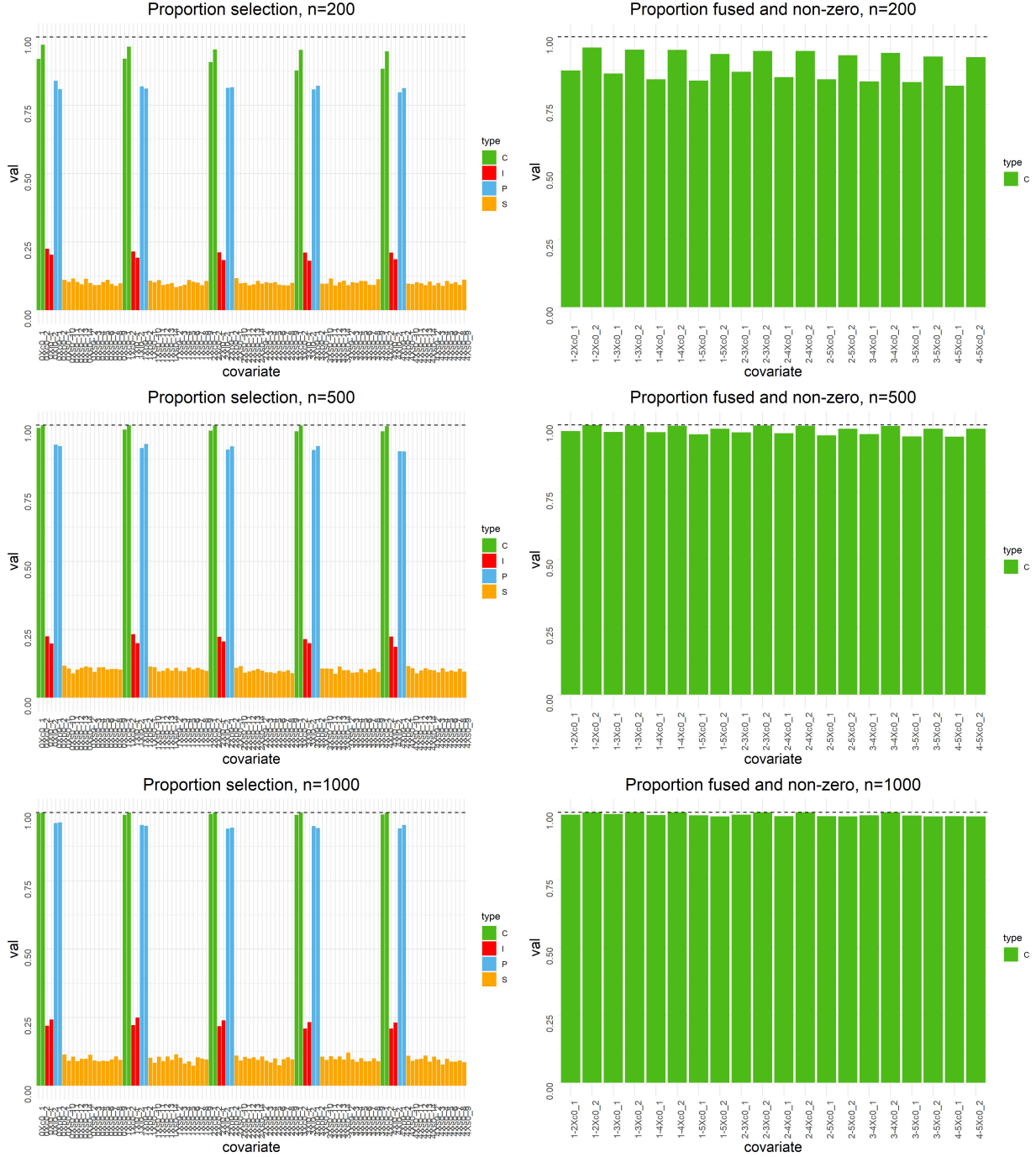| Method\Scenario	Scenario 2, two time-points, $dim(L_0)=20$, $dim(L_1)=10$			Scenario 3, five time-points, $dim(L_0)=20$		
	μ_0	μ_1	μ_2	μ_0	μ_1	μ_2
True values	1.00	0.88	0.45	0.0	1.14	0.5
n=200						
G-comp main terms	0.3(11)	0.2(5)	0.4(7)	0.1(9)	0.1(4)	0.0(1)
IPTW full main terms	3.5(118)	1.5(40)	3.7(65)	2.1(54)	0.4(16)	0.8(8)
IPTW oracle select	2.2(41)	1.6(18)	2.0(27)	0.6(22)	0.2(8)	0.2(3)
IPTW oracle select and fuse	2.2(40)	1.5(17)	2.0(27)	0.6(20)	0.1(7)	0.2(3)
LOAL	2.8(36)	1.8(17)	2.6(25)	1.1(23)	0.4(8)	0.4(3)
Fused LOAL	2.9(36)	1.7(16)	2.6(25)	1(22)	0.3(7)	0.4(3)
n=500						
G-comp main terms	0.6(9)	0.2(5)	0.8(6.5)	0.3(8)	0(3)	0.1(1)
IPTW full main terms	4.0(153)	1.7(51)	3.9(90)	2.1(73)	0.7(18)	0.9(11)
IPTW oracle select	2.1(59)	1.4(23)	2.1(41)	0.6(23)	0.3(9)	0.3(4)
IPTW oracle select and fuse	2.1(59)	1.4(23)	2.1(41)	0.6(22)	0.3(8)	0.3(3)
LOAL	3.1(45)	2.0(20)	3.0(30)	1.1(25)	0.6(9)	0.5(4)
Fused LOAL	3.2(46)	1.9(20)	3.1(31)	1(24)	0.4(9)	0.5(4)
n=1000						
G-comp main terms	0.7(9)	0.3(5)	0.8(7)	0.1(8)	0(3)	0(1)
IPTW full main terms	4.4(182)	2.5(70)	4.3(114)	1.5(80)	0.6(20)	0.7(13)
IPTW oracle select	2.8(75)	2.0(30)	2.4(48)	0.3(28)	0.3(10)	0.2(4)
IPTW oracle select and fuse	2.8(75)	2.0(30)	2.4(48)	0.3(26)	0.3(9)	0.2(4)
LOAL	3.7(59)	2.5(26)	3.3(38)	0.8(30)	0.5(10)	0.4(5)
Fused LOAL	3.7(59)	2.5(26)	3.3(39)	0.8(29)	0.3(9)	0.4(5)

Figure 4: Proportion variable selection (left) and fusion of the confounders (right) for $n = 200$ (top), $n = 500$ (middle), and $n = 1000$ (bottom) in Scenario 3. Each bar in the right-hand plots represents the proportion of fusion of each confounder at each pair of time-points.

D   Complete report on the NDIT study analysis

The Nicotine Dependence in Teens (NDIT) study is a prospective longitudinal study of 1,294 grade 7 students recruited from 10 Montréal-area (Canada) high schools in 1999-2000 O'Loughlin et al.

Table 10: $\sqrt{n}$ times absolute value of bias (n times mean squared error) for the estimation of $E(Y^{(1,1)})$ of C-LTMLE, LTMLE with propensity score covariates selected by LOAL, and LTMLE with no variable selection.

Method\Scenario	1a)	1b)	1c)	2
$E(Y^{(1,1)})$ true values	1.0	3.5	8.5	1.9
n=200				
LTMLE full	2.5(64)	10.0(348)	13.2(1907)	5.0(440)
C-LTMLE	0.3(21)	6.7(268)	10.4(1125)	0.6(18)
LTMLE LOAL	1.1(20)	5.3(208)	8.8(761)	2.6(85)
n=500				
LTMLE full	3.0(75)	13.2(635)	21.7(2770)	6.4(394)
C-LTMLE	0.3(19)	10.8(409)	17.6(1444)	1.2(23)
LTMLE LOAL	1.1(24)	9.1(348)	12.3(1185)	2.5(68)
n=1000				
LTMLE full	3.0(86)	16.7(901)	27.5(4285)	5.8(416)
C-LTMLE	1.3(21)	14.0(630)	24.7(2022)	1.9(33)
LTMLE LOAL	1.2(26)	12.0(494)	15.9(1593)	2.6(79)

[2015]. Self-report questionnaires were administered from grade 7 to 11 at each of the 10 schools every three months for a total of 20 cycles from 1999 to 2005 (i.e., during the five years of high school). Mail or in-person questionnaires were administered in 2007/2008 (cycle 21) when participants were age 20.4 years on average. The data collected include repeated measures of a wide range of socio-demographic, substance use, psychosocial, lifestyle, and physical and mental health variables.

D.1 NDIT data

D.1.1 Exposure

Participants were asked "During the past three months, how often did you drink alcohol (beer, wine, hard liquor)?" We considered a participant exposed to regular alcohol use if the participant answered "once or a couple of times a week" or "usually every day" (alternatives were "never," "a bit to try" or "once or a couple of times a month"). Therefore, "alcohol use" in this paper refers to "at least weekly use". In defining the population of interest, we excluded all participants reporting regular alcohol use at time zero. We denoted the binary exposure over time as A_t for time t.

D.1.2 Censoring

We denote the censoring indicators as C_t for each time t. A participant was censored by time t, denoted $C_t = 1$, when they were lost to follow-up or when they skipped more than one entire year of follow-up; otherwise, $C_t = 0$.

D.1.3 Covariates

Baseline covariates As baseline variables, we included socio-demographic characteristics including sex (with male as the reference sex), mothers' education (no university vs. some university), whether

the participant lived in a single-parent home, if the participant spoke French at home, and country of birth (outside Canada vs. Canada), which were assessed in the first data collection cycle. We also included: self-esteem, impulsivity, and novelty-seeking (a genetic tendency to feel intense excitement and actively explore new or potentially rewarding experiences, while also avoiding monotony and possible punishment Cloninger [1987]). While these three variables were measured in the 12-th cycle, because they are considered personality traits and unlikely to vary considerably over time, they were included as baseline covariates Liu et al. [July, 2023]. Self-esteem was measured using Rosenberg's Self-Esteem Scale Rosenberg [1965]; higher values indicate higher self-esteem Racicot et al. [2013]. Impulsivity was measured with an abbreviated version of the Eysenck Impulsivity Scale Eysenck and Eysenck [1978], which was previously validated among adolescents Wills et al. [1998]; higher scores indicate greater impulsivity Racicot et al. [2013]. Novelty-seeking was assessed using nine items based on Cloninger's Tridimensional Personality Questionnaire Otter et al. [1995]; high scores indicate greater novelty-seeking Racicot et al. [2013].

Time-varying covariates The time-varying covariates L_t were measured for time t and included: current depressive symptoms; participation in team sports; family-related stress (validated 4-point scale) with higher values indicating more stress; other type of stress (validated 4-point scale); worry about weight; and ever smoked. Unlike the outcome, current depressive symptoms were measured with a validated six-item symptoms scale Chaiton et al. [2013], Escobedo et al. [1996]; higher scores indicated higher levels of depressive symptoms. Family stress was measured using a validated scale over the past three months; higher values indicate higher levels of stress Racicot et al. [2013]. Other stress referred to the past three months with higher values indicating higher levels of stress.

D.1.4 Outcomes

The outcome Y, depression symptoms, was measured using the Major Depressive Inventory (MDI) in 2007/2008 Bech et al. [1997, 2015] . Participants asked how much time in the past two weeks they had: 1) felt low in spirit; 2) lost interest in or could no longer enjoy their daily activities; 3) felt a lack of energy and strength; 4) felt less confident; 5) had a bad conscience or feelings of guilt; 6) felt life was not worth living; 7) had difficulty concentrating; 8) felt very restless; 9) felt subdued or slowed down; 10) had trouble sleeping at night or waking up too early; 11) suffered from reduced appetite; and, 12) suffered from increased appetite. A score of four or more for items 1) and 2), and a score of three or more for the other items indicated a diagnostic demarcation for the depression symptom. For items 8) and 9), the highest score was retained for scoring, and similarly for items 11) and 12). Based on a 6-point scale ranging from 0 (*no time*) to 5 (*all the time*) for each item, responses were summed to generate a continuous score from 0 to 50 with higher scores indicating more severe symptoms Chaiton et al. [2013], Bech et al. [2015]. This scale measures depression symptoms over the past two weeks.

D.2 Handling covariate missingness in the analysis

We addressed missing values in covariates that were unrelated to censoring using imputation methods (For cases where missing data were due to censoring, we applied IPTW or LTMLE with censoring weights). To impute missing values in time-dependent covariates, we employed the Last Observation Carried Forward method which was applied for no more than one full academic year of follow-up after the last measured value. For handling missing data in baseline and the remaining time-varying covariates, we utilized multiple imputations by chained equations (MICE), mice R package Van Buuren and Groothuis-Oudshoorn [2011], maintaining time-ordering of the follow-up

variables throughout this process. Our imputation process involved a single database, and subsequent analyses were conducted to derive the estimates.

D.3 Target trial

We aim to study the effect of time of initiation of drinking during early high school on depression in young adulthood. To do so, we analyzed data from the NDIT study collected over a span of the first five cycles from 1999 to 2000 and the 21st cycle in 2007/2008. We define the target trial with corresponding intention-to-treat parameters of interest. This hypothetical trial recruits participants who have not yet initiated regular drinking at the beginning of grade 7, e.g. $A_0 = 0$ for all participants. The target trial randomizes drinking initiation to one of the follow-up time points and conducts an analysis of the correlation between time of initiation and depression symptoms in young adulthood Y.

D.4 Parameter of interest

To perform the intention-to-treat analysis with our observational data, we defined the exposure variable such that once an individual was exposed to regular alcohol use, they were considered exposed for the duration of the study unless they were lost to follow-up. Specifically, if an individual had a value of 1 for A_t at any $t = 1, \cdots, 5$, we coded the variable to set all subsequent time points, A_{t+k}, to 1 for $k = 1, \cdots, 6 - t$. Define $\mathcal{D}$ as the regimen space for the intention-to-treat analysis, then $\mathcal{D}$ represents 6 treatment patterns where the initiation time varies between 1 and 5, and no initiation for all time points, i.e.

$$
\mathcal{D} = \left\{ \begin{array}{ccccc}
(1, & 1, & 1, & 1, & 1) \\
(0, & 1, & 1, & 1, & 1) \\
(0, & 0, & 1, & 1, & 1) \\
(0, & 0, & 0, & 1, & 1) \\
(0, & 0, & 0, & 0, & 1) \\
(0, & 0, & 0, & 0, & 0)
\end{array} \right\}.
$$

Then the parameters of interest can be defined through the working marginal structural model (MSM),

$$
\mathbb{E}[Y^{\bar{a}}|\text{Sex}] = \mu_0 + \mu_1 \text{Sex} + \mu_2 \text{cum}(\bar{a}) + \mu_3 \{\text{Sex} \times \text{cum}(\bar{a})\} \tag{10}
$$

where $\mathbb{E}[Y^{\bar{a}}|\text{sex}]$ represents the mean counterfactual outcome under some intervention pattern $\bar{a}$ in a sex subgroup such that sex=1 denotes female, and $\text{cum}(\bar{a})$ counts the cumulative exposures in the pattern. The true parameter values μ minimize the expectation of a squared error loss function, summing over all patterns in the intention-to-treat space $\mathcal{D}$, corresponding to the parameters estimated in the hypothetical target trials.

D.5 Model specification

D.5.1 Data structure

Given the above, the following represents the observed data structure:

$$
O = \{\boldsymbol{L}_1, A_1, \boldsymbol{L}_2, C_2, A_2, \cdots, A_5, \boldsymbol{L}_6, C_6, Y\}.
$$

Note that $\boldsymbol{L}_1$ contains the baseline covariates and the time-varying covariates at the first time and there is no censoring prior to the first exposure time.

D.5.2 Outcome models

We use the notation $\bar{\boldsymbol{L}}_t$ to denote the history of baseline and time-dependent covariates up to time t and likewise $\bar{\boldsymbol{A}}_t$ represents the history of the exposure $A_1, \cdots, A_t$. We rescaled the bounded continuous outcome Y to be contained in $(0, 1)$. Denote $T = 6$ as the total number of time points. Starting with $q_{T+1}(\bar{\boldsymbol{a}}_{T+1}, \bar{\boldsymbol{L}}_{T+1}) = Y$, we recursively define

$$q_t(\bar{\boldsymbol{a}}_t, \bar{\boldsymbol{L}}_t) = \mathbb{E}\{q_{t+1}(\bar{\boldsymbol{a}}_{t+1}, \bar{\boldsymbol{L}}_{t+1})|\bar{\boldsymbol{L}}_t, C_t = 0, \bar{\boldsymbol{A}}_t = \bar{\boldsymbol{a}}_t\}, \quad t = T, \cdots, 1.$$

For the NDIT data setting, the history of the exposures up to time T is equal to the history of the exposures to time $T - 1$. In order to obtain preliminary estimates $q_{t,n}(\bar{\boldsymbol{a}}_t, \bar{\boldsymbol{L}}_t)$ of $q_t(\bar{\boldsymbol{a}}_t, \bar{\boldsymbol{L}}_t)$, we modeled the outcome for $t = T$ or most recent estimate of $q_{t+1}(\bar{\boldsymbol{a}}_t, \bar{\boldsymbol{L}}_t)$, conditional on main terms of the baseline and time varying covariates, exposure terms (current and lagged) and the first-order interactions of sex and exposure terms for uncensored participants. Then we generated predictions from this model for each pattern of interest $\bar{\boldsymbol{a}}$. We fit logistic regressions stratified on time t, corresponding to:
For $t = T$,

$$Y \sim \sum_{k=1}^{t} \boldsymbol{L}_k + \sum_{k=1}^{t-1} A_k + \sum_{k=1}^{t-1} \{\text{Sex} \times A_k\}$$

For $t = T - 1, \cdots, 1$,

$$q_{t+1,n}(\bar{\boldsymbol{a}}) \sim \sum_{k=1}^{t} \boldsymbol{L}_k + \sum_{k=1}^{t} A_k + \sum_{k=1}^{t} \{\text{Sex} \times A_k\}.$$

Taking as outcome a vector composed of stacked components $q_{t,n}(\bar{\boldsymbol{a}}_t, \bar{\boldsymbol{L}}_t)$ for each pattern $\bar{\boldsymbol{a}}$, we then run regressions according to the following working regression models:

$$\hat{E}\{q_{1,n} \mid \bar{\boldsymbol{L}}_1, a_1\} = \beta_{1,0} + \boldsymbol{\beta}_{1,1}\boldsymbol{L}_1 + \beta_{1,2}a_1,$$
$$\hat{E}\{q_{2,n} \mid \bar{\boldsymbol{L}}_2, C_2 = 0, \bar{\boldsymbol{a}}_2\} = \beta_{2,0} + \boldsymbol{\beta}_{2,1}\boldsymbol{L}_1 + \boldsymbol{\beta}_{2,2}\boldsymbol{L}_2 + \beta_{2,3}a_1 + \beta_{2,4}a_2,$$
$$\vdots$$
$$\hat{E}\{q_{T,n} \mid \bar{\boldsymbol{L}}_T, C_T = 0, \bar{\boldsymbol{a}}_T\} = \beta_{T,0} + \sum_{k=1}^{T}\boldsymbol{\beta}_{T,k}\boldsymbol{L}_k + \sum_{k=1}^{T}\beta_{T,T+k}a_k$$

$$\tag{11}$$

with true parameter values $\boldsymbol{\beta} = \{\beta_{\tau,t}; \tau = (1, \cdots, T),\ t = (0, \cdots, 2\tau)\}$ minimizing the risk under a squared-error loss function.

D.5.3 Pooled treatment model and pooled censoring model

As discussed in the manuscript, we define a "full" model for the probability of treatment at all time-points that adjusts for the full covariate history. The treatment model was fit using those who had not yet initiated and were uncensored at each time point. A pooled logistic regression for the conditional probability of treatment, $P(A_t = 1 \mid A_{t-1} = 0, \bar{\boldsymbol{L}}_t, C_t = 0)$, was specified as follows. We

define $m_t(\bar{\boldsymbol{L}}_t; \boldsymbol{\alpha})$ as the corresponding model for the probability of treatment at times $t = 1, \cdots, T$.

$$\text{logit}\left\{m_t(\bar{\boldsymbol{L}}_t; \boldsymbol{\alpha})\right\}$$
$$= \mathbb{I}(t = 1)\left(\alpha_{1,0} + \boldsymbol{\alpha}_{1,1}\boldsymbol{L}_1\right) +$$
$$\mathbb{I}(t = 2)\left(\alpha_{2,0} + \boldsymbol{\alpha}_{2,1}\boldsymbol{L}_1 + \boldsymbol{\alpha}_{2,2}\boldsymbol{L}_2\right) +$$
$$\vdots$$
$$\mathbb{I}(t = T - 1)\left(\alpha_{T-1,0} + \sum_{k=1}^{T-1}\boldsymbol{\alpha}_{T-1,k}\boldsymbol{L}_k\right).$$

In the above, $\boldsymbol{\alpha} = \{\alpha_{\tau,t}; \tau = (1, \cdots, T - 1),\ t = (0, \cdots, \tau)\}$ are the coefficients of the covariates in the pooled propensity score model. Notably, the exposure does not appear in this model since it is fit on the subset of participants who had not yet initiated drinking (and so all past exposure is null).

In addition, the pooled censoring model adjusted for the history of covariates and treatments in order to estimate $h_t(\bar{\boldsymbol{L}}_t, \bar{A}_{t-1}; \boldsymbol{\theta}) = P(C_t = 0 \mid \bar{\boldsymbol{L}}_t, \bar{A}_{t-1}, C_{t-1} = 0); t = 2, \cdots, T$. The model was specified as

$$\text{logit}\left\{1 - h_t(\bar{\boldsymbol{L}}_t, \bar{A}_{t-1},; \boldsymbol{\theta})\right\}$$
$$= \mathbb{I}(t = 2)\left(\theta_{2,0} + \boldsymbol{\theta}_{2,1}\boldsymbol{L}_1 + \boldsymbol{\theta}_{2,2}\boldsymbol{L}_2 + \theta_{2,3}A_1\right) +$$
$$\vdots$$
$$\mathbb{I}(t = T)\left(\theta_{T,0} + \sum_{k=1}^{T}\boldsymbol{\theta}_{T,k}\boldsymbol{L}_k + \sum_{k=1}^{T-1}\theta_{T,T+k}A_k\right).$$

In the above, $\boldsymbol{\theta} = \{\theta_{\tau,t}; \tau = (2, \cdots, T),\ t = (0, \cdots, 2\tau - 1)\}$ are the coefficients of the covariates in the pooled censoring model.

D.6   Cumulative weights for treatment and censoring

To define the weights used in estimation and the balance criteria, we must extend our definition of the model for the probability of exposure to be deterministic when exposure was already initiated in the past. Thus, we define

$$m_t^*(\bar{\boldsymbol{L}}_t, a_{t-1}; \boldsymbol{\alpha}) = \begin{cases} m_t(\bar{\boldsymbol{L}}_t; \boldsymbol{\alpha}) & for\ a_{t-1} = 0 \\ 1 & for\ a_{t-1} = 1. \end{cases}$$

The cumulative weights for treatment at times $t = (1, \cdots, T - 1)$ are

$$w_t^a(a_t, \bar{\boldsymbol{L}}_t; \boldsymbol{\alpha}) = \frac{\mathbb{I}(A_t = a_t)}{a_t m_t(\bar{\boldsymbol{L}}_t; \boldsymbol{\alpha}) + (1 - a_t)[1 - m_t(\bar{\boldsymbol{L}}_t; \boldsymbol{\alpha})]} \quad \text{for } t = 1, \text{and}$$

$$w_t^a(a_t, \bar{a}_{t-1}, \bar{\boldsymbol{L}}_t; \boldsymbol{\alpha}) = \frac{\mathbb{I}(A_t = a_t, \bar{A}_{t-1} = \bar{a}_{t-1})}{\prod_{k=1}^{t} a_k m_k^*(a_{k-1}, \bar{\boldsymbol{L}}_k; \boldsymbol{\alpha}) + (1 - a_k)[1 - m_k^*(a_{k-1}, \bar{\boldsymbol{L}}_k; \boldsymbol{\alpha})]} \quad \text{for } t = 2, \cdots, T - 1. \tag{12}$$

The cumulative weight for censoring at time $t = (2, \cdots, T)$ is

$$w_t^c(\bar{\boldsymbol{L}}_t, \bar{a}_{t-1}; \boldsymbol{\theta}) = \prod_{k=2}^{t} \frac{\mathbb{I}(C_k = 0)}{h_k(\bar{\boldsymbol{L}}_k, \bar{a}_{k-1}; \boldsymbol{\theta})}. \tag{13}$$

S13

Combinations of these weights are used in the balance criteria.

The cumulative weights used in IPTW are written as: $w_t^{iptw} = 1/P(A_t = a_t \mid \boldsymbol{L}_t = \boldsymbol{l}_t)$ for time $t = 1$ and $w_t^{iptw} = 1/P(A_1 = a_1 \mid \boldsymbol{L}_1 = \boldsymbol{l}_1) \prod_{k=2}^{t} w_k$ for time $t = (2, \cdots, T)$ where

$$
w_k =
\begin{cases}
\dfrac{\mathbb{I}(C_k = 0)}{P(A_k = a_k \mid a_{k-1} = 0, \bar{\boldsymbol{L}}_k = \bar{\boldsymbol{l}}_k, C_k = 0)P(C_k = 0 \mid \bar{\boldsymbol{L}}_k = \bar{\boldsymbol{l}}_k, C_{k-1} = 0, a_{k-1} = 0)} & for\ a_{k-1} = 0 \\[4mm]
\dfrac{\mathbb{I}(C_k = 0)}{P(C_k = 0 \mid \bar{\boldsymbol{L}}_k = \bar{\boldsymbol{l}}_k, C_{k-1} = 0, \bar{\boldsymbol{A}}_{k-1} = \bar{\boldsymbol{a}}_{k-1})} & for\ a_{k-1} = 1.
\end{cases}
$$

D.7 Longitudinal Outcome Adaptive Lasso

We implemented LOAL to select variables for the treatment models and censoring models separately at each time that have corresponding non-zero coefficients $\boldsymbol{\beta}$ in the q_t model fits.

D.7.1 LOAL for treatment

Given a regularization parameter $\lambda_n^a \geq 0$, the pooled LOAL estimator for $\boldsymbol{\alpha}^{\dagger}$ is given as,

$$
\hat{\boldsymbol{\alpha}}(\lambda_n^a) = \arg\min_{\alpha} \sum_{\tau=1}^{T} \sum_{i=1}^{n} \big[a_{\tau,i} \log\{m_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\alpha})\}
$$
$$
+ (1 - a_{\tau,i}) \log\{1 - m_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\alpha})\}\big] + \lambda_n^a \sum_{j \in \mathcal{J}^a} \hat{\omega}_j |\alpha_j|,
$$

where $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma}$ for $j \in \mathcal{J}^a$ and $\gamma = 2.5$. Here $\mathcal{J}^a$ represents the indices of the coefficients $\boldsymbol{\alpha}$ being shrunk, i.e.,

$$
\begin{aligned}
\mathcal{J}^a = \{ &(1, 1, \mathcal{J}_{1,1}^a), \\
&(2, 1, \mathcal{J}_{2,1}^a), (2, 2, \mathcal{J}_{2,2}^a), \\
&\vdots \\
&(T - 1, 1, \mathcal{J}_{T-1,1}^a), (T - 1, 2, \mathcal{J}_{T-1,2}^a), \cdots, (T - 1, T - 1, \mathcal{J}_{T-1,T-1}^a)\}
\end{aligned}
$$

where $\mathcal{J}_{\tau,t}^a$, for $\tau, t = (1, \cdots, T - 1), t \leq \tau$, represents the indices of set of coefficients at each time. Specifically, $\mathcal{J}_{\tau,t}^a$ indexes the specific covariates in $\boldsymbol{L}_t$ being shrunk within propensity score model $\boldsymbol{A}_\tau$.

This regularized regression for the treatment can be implemented by a transformation of the pooled data, setting

$$
\begin{aligned}
&V_{1,0} = \mathbb{I}(t = 1), \boldsymbol{V}_{1,1} = \mathbb{I}(t = 1)\boldsymbol{L}_1, \\
&V_{2,0} = \mathbb{I}(t = 2), \boldsymbol{V}_{2,1} = \mathbb{I}(t = 2)\boldsymbol{L}_1, \boldsymbol{V}_{2,2} = \mathbb{I}(t = 2)\boldsymbol{L}_2, \\
&\vdots \\
&V_{T-1,0} = \mathbb{I}(t = T - 1), \boldsymbol{V}_{T-1,1} = \mathbb{I}(t = T - 1)\boldsymbol{L}_1, \cdots, \boldsymbol{V}_{T-1,T-1} = \mathbb{I}(t = T - 1)\boldsymbol{L}_{T-1}
\end{aligned}
$$

with respectively corresponding coefficients in $\boldsymbol{\alpha}$. Then, the adaptive LASSO is run with pooled outcome A_τ on these covariates $\boldsymbol{V}_{1,0}, ..., \boldsymbol{V}_{T-1,T-1}$, without an intercept term, using weights $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma}; j \in \mathcal{J}^a$.

D.7.2 LOAL for censoring

For the censoring, given the same $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma}$ for $j \in \mathcal{J}^c$ and $\gamma = 2.5$, the pooled LOAL estimator $\boldsymbol{\theta}_0^\dagger$ on $\lambda_n^c \geq 0$ is,

$$\hat{\boldsymbol{\theta}}(\lambda_n^c) = \arg\min_{\theta} \sum_{\tau=2}^{T} \sum_{i=1}^{n} \big[(1 - c_{\tau,i}) \log\{h_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\theta})\}$$

$$+ c_{\tau,i} \log\{1 - h_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\theta})\} \big] + \lambda_n^c \sum_{j \in \mathcal{J}^c} \hat{\omega}_j |\theta_j|,$$

where $\mathcal{J}^c$ represents the indices of the coefficients $\boldsymbol{\theta}$ being shrunk. $\mathcal{J}_{\tau,t}^c$ for $\tau = (2, \cdots, T), t = (1, \cdots, T), t \leq \tau$ indexes the specific covariates in $\boldsymbol{L}_t$ within the censoring models C_τ. Note that the intercepts and the coefficients corresponding to treatments are not shrunk,

$$\mathcal{J}^c = \{(2, 1, \mathcal{J}_{2,1}^c), (2, 2, \mathcal{J}_{2,2}^c),$$
$$(3, 1, \mathcal{J}_{3,1}^c), (3, 2, \mathcal{J}_{3,2}^c), (3, 3, \mathcal{J}_{3,3}^c),$$
$$\vdots$$
$$(T, 1, \mathcal{J}_{T,1}^c), (T, 2, \mathcal{J}_{T,2}^c), \cdots, (T, T, \mathcal{J}_{T,T}^c)\}.$$

Likewise, this regularization for the censoring can be applied through a transformation of the pooled data, but encompassing not only the variables of time and covariates but also treatment related variables,

$$U_{2,0} = \mathbb{I}(t = 2), U_{2,1} = \mathbb{I}(t = 2)\boldsymbol{L}_1, U_{2,2} = \mathbb{I}(t = 2)\boldsymbol{L}_2, U_{2,3} = \mathbb{I}(t = 2)A_1,$$
$$\vdots$$
$$U_{T,0} = \mathbb{I}(t = T), U_{T,1} = \mathbb{I}(t = T)\boldsymbol{L}_1, \cdots, U_{T,T} = \mathbb{I}(t = T)\boldsymbol{L}_T,$$
$$U_{T,T+1} = \mathbb{I}(t = T)A_1, \cdots, U_{T,2T-1} = \mathbb{I}(t = T)A_{T-1}$$

with respectively corresponding coefficients in $\boldsymbol{\theta}$. Then, the adaptive LASSO is run with pooled outcome C_t on these covariates $U_{2,0}, ..., U_{T,2T-1}$ without an intercept term.

D.8 Selection of λ_n^a and λ_n^c

The tuning parameter λ_n^a and λ_n^c were selected jointly by minimizing the sum of the balance metrics for treatment and censoring. The balance metric for treatment, $\mathcal{M}$, is a summary of weighted absolute mean differences (wAMDs) of the covariates between the exposure groups Shortreed and Ertefaie [2017]. Similarly, the balance metric for censoring, $\mathcal{N}$, is based on the wAMDs of the covariates between uncensored and censored groups.

The weights involved are cumulative inverse probability weights for current treatment or censoring. For balance across treatment groups at a given time, we only consider histories with no past exposure, since the only comparison to make is in people who initiated or did not initiate exposure. Let $\hat{\boldsymbol{\alpha}}^{refit}(\lambda_n^a)$ represent the estimates from a logistic regression of the treatment on the covariates selected by LOAL under tuning parameter λ_n^a where the value is set to be zero if the corresponding coefficient was not selected. Similarly, $\hat{\boldsymbol{\theta}}^{refit}(\lambda_n^c)$ represent the estimates from a logistic regression of the censoring on the covariates selected by LOAL under tuning parameter λ_n^c. Based on equations (12) and (13), we define the weight for subject i at time t for the current treatment as

$$w_{t,i}^a = w_t^a\{a_{t,i}, \bar{\boldsymbol{a}}_{t-1,i} = 0, \bar{\boldsymbol{l}}_{t,i}; \hat{\boldsymbol{\alpha}}^{refit}(\lambda_n^a)\} w_t^c\{\bar{\boldsymbol{a}}_{t-1,i}, \bar{\boldsymbol{l}}_{t,i}; \hat{\boldsymbol{\theta}}^{refit}(\lambda_n^c)\}$$

where $\hat{\boldsymbol{\alpha}}(\lambda_n^a)$ and $\hat{\boldsymbol{\theta}}(\lambda_n^c)$ are parameter estimates under the treatment and censoring model after variable selection by LOAL with the tuning parameters $(\lambda_n^a, \lambda_n^c)$, respectively. Also,

$$w_{t,i}^c = w_t^c\{\bar{\boldsymbol{a}}_{t-1,i}, \bar{\boldsymbol{l}}_{t,i}; \hat{\boldsymbol{\theta}}^{refit}(\lambda_n^c)\}w_{t-1}^a\{a_{t-1,i}, \bar{\boldsymbol{l}}_{t-1,i}; \hat{\boldsymbol{\alpha}}^{refit}(\lambda_n^a)\}$$

is the weight for current censoring for subject i at time t estimated under the proposed LOAL approach of censoring and treatment. Let $L_{t,k}$ denote the kth component in $\boldsymbol{L}_t$ for $k = (1, \cdots, p_k)$ where p_k represents the number of components of $\boldsymbol{L}_t$. Then $\beta_{\tau,t,k}$ represents the coefficients in the structural equations (equations 11) for $\tau = (1, \cdots, T-1)$ and $\tau = (2, \cdots, T)$ referring to the treatment model and censoring model respectively. Then the weighted absolute mean difference of the treatment and of the censoring can be evaluated based on the variable considered at time t respectively weighted by the corresponding structural models coefficient (equations 11) divided by its standard error.

$\text{wAMD}_{\tau,t,k}^a =$

$$\frac{\mid \hat{\beta}_{\tau,t,k} \mid}{\sigma_{\hat{\beta}_{\tau,t,k}}} \left| \frac{\sum_{i=1}^n a_{\tau,i} l_{t,k,i} w_{\tau,i}^a \mathbb{I}(a_{\tau-1,i} = 0, c_{\tau,i} = 0)}{\sum_{i=1}^n a_{\tau,i} w_{\tau,i}^a \mathbb{I}(a_{\tau-1,i} = 0, c_{\tau,i} = 0)} - \frac{\sum_{i=1}^n (1 - a_{\tau,i}) l_{t,k,i} w_{\tau,i}^a \mathbb{I}(a_{\tau-1,i} = 0, c_{\tau,i} = 0)}{\sum_{i=1}^n (1 - a_{\tau,i}) w_{\tau,i}^a \mathbb{I}(a_{\tau-1,i} = 0, c_{\tau,i} = 0)} \right|$$

for $\tau = (1, \cdots, T-1), t = (1, \cdots, T-1)$ and $t \le \tau$.

$$\text{wAMD}_{\tau,t,k}^c = \frac{\mid \hat{\beta}_{\tau,t,k} \mid}{\sigma_{\hat{\beta}_{\tau,t,k}}} \left| \frac{\sum_{i=1}^n \mathbb{I}(c_{\tau,i} = 0) l_{t,k,i} w_{\tau,i}^c}{\sum_{i=1}^n \mathbb{I}(c_{\tau,i} = 0) w_{\tau,i}^c} - \frac{\sum_{i=1}^n \{\mathbb{I}(c_{\tau,i} = 1) l_{t,k,i} w_{\tau,i}^c\}}{\sum_{i=1}^n \{\mathbb{I}(c_{\tau,i} = 1) w_{\tau,i}^c\}} \right|$$

for $\tau = (2, \cdots, T), t = (1, \cdots, T)$ and $t \le \tau$

We selected the two tuning parameters by minimizing the sum of the balance criterion for the treatment and the balance criterion for the censoring, i.e. the selected $(\lambda_n^a, \lambda_n^c) = \arg\min_{\lambda_n^a, \lambda_n^c} (\mathcal{M} + \mathcal{N})$ where

$$\mathcal{M} = \sum_{k=1}^{p_1} \text{wAMD}_{1,1,k}^a + \sum_{t=1}^2 \sum_{k=1}^{p_k} \text{wAMD}_{2,t,k}^a + \cdots + \sum_{t=1}^{T-1} \sum_{k=1}^{p_k} \text{wAMD}_{T-1,t,k}^a$$

$$= \sum_{\tau=1}^{T-1} \sum_{t=1}^{\tau} \sum_{k=1}^{p_k} \text{wAMD}_{\tau,t,k}^a$$

$$\mathcal{N} = \sum_{t=1}^2 \sum_{k=1}^{p_k} \text{wAMD}_{2,t,k}^c + \sum_{t=1}^3 \sum_{k=1}^{p_k} \text{wAMD}_{3,t,k}^c + \cdots + \sum_{t=1}^T \sum_{k=1}^{p_k} \text{wAMD}_{T,t,k}^c$$

$$= \sum_{\tau=2}^T \sum_{t=1}^{\tau} \sum_{k=1}^{p_k} \text{wAMD}_{\tau,t,k}^c$$

D.9 Selective fusion

To perform the selective fusion, we initially establish a penalty graph, wherein vertices represent coefficients within the pooled model eligible for fusion. This graph may be structured with cliques connecting elements that share common variable names across different time points. For example, it links the remaining baseline covariates in different treatment model times. As for time-varying

covariates, we created cliques to allow coefficient fusion of the most recent variables of common types across various time points; this allows for the possibility of common effects of historical covariates with common lag time on current exposure initiation. Specifically, the penalty graph connected the same baseline variables across time points, and the same time-varying variables with the same lag across time points (e.g., the $\boldsymbol{L}_{t-1}$ variables are connected when modeling treatment and censoring across times t).

For example, consider the propensity score model for A_4 and a particular time varying covariate $L_{t,3}$. Suppose that both $L_{3,3}$ (the most recent) and $L_{4,3}$ (the current) were selected into this model. In the model for A_5, suppose that $L_{4,3}$ and $L_{5,3}$ were selected. In the model for A_6, suppose that $L_{4,3}$ and $L_{6,3}$ were selected. The cliques connect the coefficients of the three current variables, $L_{4,3}, L_{5,3}$, and $L_{6,3}$ across the models for A_4, A_5 and A_6, respectively, and also connect the two most recent variables, $L_{3,3}$ and $L_{4,3}$ in the models for A_4 and A_5, respectively.

Denote $\mathcal{G}$ as the set of all pairwise connected indices of the coefficients in accordance with the fusion graph definition. The fused LASSO penalizes the absolute differences between the coefficients of connected variables.

Define $\boldsymbol{\alpha}^*$ as the parameter vector of the same length as $\boldsymbol{\alpha}$ that is set to zero at the indices of the zero-elements of $\hat{\boldsymbol{\alpha}}^{refit}(\lambda_n^a)$. Then the generalized Adaptive Fused LASSO for treatments is

$$
\arg\min_{\boldsymbol{\alpha}^*} \sum_{\tau=1}^{T-1} \sum_{i=1}^{n} \left[a_{\tau,i} \log\{m_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\alpha}^*)\} + (1 - a_{\tau,i}) \log\{1 - m_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\alpha}^*)\} \right]
$$
$$
+ \lambda_{1,n}^a \sum_{(\mathcal{J}_{\tau,t}^a, \mathcal{J}_{\tau',t'}^a) \in \mathcal{G}^a} \frac{|\boldsymbol{\alpha}_{\mathcal{J}_{\tau,t}^a}^* - \boldsymbol{\alpha}_{\mathcal{J}_{\tau',t'}^a}^*|}{|\hat{\boldsymbol{\alpha}}_{\mathcal{J}_{\tau,t}^a}^{refit}(\lambda_n^a) - \hat{\boldsymbol{\alpha}}_{\mathcal{J}_{\tau',t'}^a}^{refit}(\lambda_n^a)|^\tau}.
$$

where $(\mathcal{J}_{\tau,t}^a, \mathcal{J}_{\tau',t'}^a)$ is a pair of indices in the graph for treatment $\mathcal{G}^a$. Note that τ, τ', t, t' all in $(1, \cdots, T-1)$, and $t \leq \tau, t' \leq \tau', \tau \neq \tau', t \neq t'$, and $\tau - \tau' = t - t'$ based on the penalty graph we defined.

The penalty graph for censoring $\mathcal{G}^c$ was created in the same way as the graph for treatment model. Note that all coefficients corresponding to the treatment terms were not allowed to fuse. Denote $\boldsymbol{\theta}^*$ as the vector of the same length as $\boldsymbol{\theta}$ that is set to zero at the indices of the zero-elements of $\hat{\boldsymbol{\theta}}(\lambda_n^c)$. The generalized Adaptive Fused LASSO for the censoring is

$$
\arg\min_{\boldsymbol{\theta}^*} \sum_{\tau=2}^{T} \sum_{i=1}^{n} \left[(1 - c_{\tau,i}) \log\{h_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\theta}^*)\} + c_{\tau,i} \log\{1 - h_\tau(\bar{\boldsymbol{l}}_{\tau,i}, \bar{\boldsymbol{a}}_{\tau-1,i}; \boldsymbol{\theta}^*)\} \right]
$$
$$
+ \lambda_{1,n}^c \sum_{(\mathcal{J}_{\tau,t}^c, \mathcal{J}_{\tau',t'}^c) \in \mathcal{G}^c} \frac{|\boldsymbol{\theta}_{\mathcal{J}_{\tau,t}^c}^* - \boldsymbol{\theta}_{\mathcal{J}_{\tau',t'}^c}^*|}{|\hat{\boldsymbol{\theta}}_{\mathcal{J}_{\tau,t}^c}^{refit}(\lambda_n^c) - \hat{\boldsymbol{\theta}}_{\mathcal{J}_{\tau',t'}^c}^{refit}(\lambda_n^c)|^\tau},
$$

where $(\mathcal{J}_{\tau,t}^c, \mathcal{J}_{\tau',t'}^c)$ is a pair of connected indices in the graph for censoring $\mathcal{G}^c$ and τ, τ', t, t' all in $(2, \cdots, T)$, and $t \leq \tau, t' \leq \tau', \tau \neq \tau', t \neq t'$, and $\tau - \tau' = t - t'$. We utilized the archived `FusedLasso` package to implement the fusion step. The selection of optimal $\lambda_{1,n}^a$ and $\lambda_{1,n}^c$ values for the treatment and censoring models was determined based on the summation of the Bayesian Information Criterion (BIC) of the treatment and censoring models.

D.10 NDIT results

The intention-to-treat analysis of the NDIT data included eight baseline covariates and six time-varying covariates. The pooled treatment model included 130 variables, while the pooled censoring

model conditioned on 175 variables. To regularize our models, we set the tuning parameter γ to 2.5 and set 20 possible values for the tuning parameters λ^a and λ^c (refer to Table 11). The selected tuning parameters, $[\lambda_n^a, \lambda_n^c]$, were found to be $[3.728, 67.392]$. These values corresponded to a strong penalty for the treatment model and a relatively light penalty for the censoring model. For the fusion step, we considered 20 possible values for $\lambda_{1,n}^a$ within the range $[e^{-10}, e^{-1}]$ and for $\lambda_{1,n}^c$ within the range $[e^{-5}, 1]$. Ultimately, the minimum summation of BICs corresponding to the treatment and censoring was achieved with $\lambda_{1,n}^a$ in $[0.002, 0.368]$ and $\lambda_{1,n}^c$ in $[0.206, 1]$.

The initial treatment model, with 135 parameters (including five intercepts), was reduced to 37 parameters. Consequently, the fusion step further reduced the number of parameters to 23 (see Table in the main manuscript). The variables sex, country of birth, current depressive symptoms, and worry about weight were selected for inclusion in each time period, and their corresponding coefficients were then fused. The variables mother education, ever smoked, family stress, other stress, and team sports were selected into the models for some time-points but were not fused. For the censoring models, initially, there were 180 parameters (including five intercepts and 15 past treatments), which were later reduced to 112 due to selection and further fused to produce a total of 55 parameters. The selected and fused variables in this case included sex, country of birth, current depressive symptoms, ever smoked, family-related stress, other stress, participation in team sports and worry about weight (Table with results is in the main manuscript).

We also applied LTMLE to estimate the target parameters of the MSM. Based on the same outcome models which involved all covariate main terms and the interaction terms between sex and treatments, we implemented: 1) LTMLE using propensity scores on the full set of covariates; 2) LTMLE using propensity scores on the selected set of covariates by LOAL; 3) LTMLE using propensity scores after LOAL selection and fusion. Estimated coefficients and standard errors are presented in the Table in the main manuscript. We used the `sandwich` R package to estimate the robust sandwich variance of IPTW and the variance of LTMLE was estimated based on influence function. All methods consistently demonstrated that being female was associated with more severe depressive symptoms when compared to males, who served as the reference group. IPTW full and LTMLE full had point-estimates that suggested that early alcohol initiation was linked to detrimental effects on depressive symptoms in males. All IPTW point-estimates suggested that earlier alcohol initiation was beneficial for females, while the LTMLE results indicated harmful or null effects for females. Furthermore, propensity scores derived from covariates selected by LOAL and then fused led to apparent reduced estimation variance in both IPTW and LTMLE analyses.

Table 11: Combinations for balance criteria based on 20 values of λ^a and the first row represents 20 values of λ^c in the selective step. In each cell, the left number represents the balance criterion of treatment $\mathcal{M}$ for the given λ^a, and the right number represents the balance criterion of censoring $\mathcal{N}$ for the given λ^c. The pair of numbers highlighted was selected to minimize the summation of the balance criteria.

λ_n^a/λ_n^c	1 (22026.466)	2 (10001.86)	3 (4541.682)	4 (2062.304)	5 (936.459)	6 (425.231)	7 (193.09)	8 (87.679)	9 (39.814)	10 (18.079)
2980.958	(16.369, 46.532)	(16.369, 46.532)	(16.31, 46.578)	(16.377, 47.207)	(16.306, 48.167)	(16.292, 48.351)	(16.256, 48.497)	(16.218, 47.631)	(16.242, 47.788)	(16.181, 48.484)
1585.129	(14.432, 49.587)	(14.432, 49.336)	(14.389, 49.472)	(14.455, 48.651)	(14.334, 48.255)	(14.313, 48.281)	(14.272, 48.263)	(14.202, 47.601)	(14.196, 47.703)	(14.172, 47.805)
842.895	(13.629, 51.104)	(13.629, 51.104)	(13.593, 51.027)	(13.632, 50.253)	(13.504, 50.003)	(13.484, 50.035)	(13.45, 50.08)	(13.389, 49.672)	(13.398, 49.559)	(13.408, 49.453)
448.211	(13.517, 51.445)	(13.517, 51.445)	(13.482, 51.374)	(13.521, 50.695)	(13.394, 50.409)	(13.366, 50.433)	(13.331, 50.577)	(13.283, 50.157)	(13.294, 49.956)	(13.325, 49.929)
238.337	(12.925, 49.895)	(12.925, 49.895)	(12.886, 49.772)	(12.928, 49.003)	(12.882, 48.973)	(12.882, 49.058)	(12.85, 49.057)	(12.78, 48.514)	(12.734, 48.773)	(12.691, 48.791)
126.736	(13.011, 49.078)	(13.011, 49.078)	(12.974, 49.043)	(13.024, 49.144)	(12.987, 52.875)	(12.988, 53.166)	(12.977, 53.071)	(12.916, 52.259)	(12.863, 52.665)	(12.793, 53.189)
67.392	(13.233, 48.36)	(13.233, 48.36)	(13.196, 48.367)	(13.25, 47.889)	(13.138, 48.241)	(13.142, 48.33)	(13.128, 48.433)	(13.102, 47.836)	(13.078, 47.709)	(13.001, 47.548)
35.836	(13.769, 48.901)	(13.769, 48.901)	(13.729, 48.709)	(13.782, 48.709)	(13.8, 53.452)	(13.889, 53.535)	(13.809, 53.453)	(13.809, 52.908)	(13.757, 52.433)	(13.534, 53.065)
19.056	(12.23, 63.313)	(12.23, 63.313)	(12.182, 63.526)	(12.242, 65.978)	(12.221, 70.768)	(12.281, 70.792)	(12.3, 70.589)	(12.196, 69.671)	(12.147, 69.004)	(12, 69.429)
10.133	(14.396, 79.943)	(14.396, 79.943)	(14.352, 80.08)	(14.425, 81.565)	(14.339, 81.133)	(14.36, 81.155)	(14.372, 81.368)	(14.318, 82.405)	(14.272, 81.373)	(14.184, 82.627)
5.388	(13.033, 50.673)	(13.033, 50.673)	(12.976, 50.8)	(13.07, 51.984)	(12.953, 56.357)	(12.953, 56.32)	(12.931, 55.807)	(12.924, 54.88)	(12.896, 55.083)	(12.845, 55.337)
2.865	(11.737, 78.484)	(11.737, 78.484)	(11.681, 78.818)	(11.779, 78.971)	(11.66, 79.397)	(11.653, 79.338)	(11.63, 79.895)	(11.645, 79.97)	(11.624, 78.424)	(11.582, 79.129)
1.524	(10.21, 73.849)	(10.21, 73.849)	(10.159, 74.314)	(10.256, 74.081)	(10.101, 75.355)	(10.085, 75.27)	(10.057, 75.701)	(10.091, 76.048)	(10.08, 74.789)	(10.075, 75.687)
0.81	(10.125, 78.234)	(10.125, 78.234)	(10.069, 78.646)	(10.166, 77.191)	(10.069, 78.387)	(10.075, 78.158)	(10.062, 78.571)	(10.056, 79.251)	(10.035, 78.483)	(9.979, 78.849)
0.431	(9.926, 77.562)	(9.926, 77.562)	(9.872, 77.969)	(9.964, 76.484)	(9.85, 77.671)	(9.848, 77.439)	(9.835, 77.861)	(9.84, 78.573)	(9.815, 77.751)	(9.784, 78.117)
0.229	(10.198, 82.721)	(10.198, 82.721)	(10.137, 82.982)	(10.232, 82.354)	(10.147, 82.951)	(10.145, 82.762)	(10.127, 82.964)	(10.128, 83.23)	(10.105, 83.2)	(10.051, 83.456)
0.122	(9.681, 85.639)	(9.681, 85.639)	(9.626, 85.92)	(9.724, 85.554)	(9.629, 85.726)	(9.614, 85.573)	(9.587, 85.704)	(9.586, 85.678)	(9.571, 86.078)	(9.547, 86.678)
0.065	(10.333, 85.462)	(10.333, 85.462)	(10.278, 85.721)	(10.376, 85.195)	(10.271, 85.544)	(10.264, 85.396)	(10.238, 85.512)	(10.25, 85.467)	(10.238, 85.73)	(10.223, 86.233)
0.034	(10.376, 84.707)	(10.376, 84.707)	(10.32, 85.021)	(10.414, 84.404)	(10.314, 84.796)	(10.307, 84.648)	(10.285, 84.789)	(10.296, 84.762)	(10.288, 85.052)	(10.265, 85.627)
0.018	(9.842, 82.699)	(9.842, 82.699)	(9.786, 83.157)	(9.878, 82.589)	(9.775, 82.927)	(9.772, 82.786)	(9.763, 82.662)	(9.754, 82.957)	(9.747, 83.58)	(9.702, 84.764)

λ_n^a/λ_n^c	11 (8.209)	12 (3.728)	13 (1.693)	14 (0.769)	15 (0.349)	16 (0.158)	17 (0.072)	18 (0.033)	19 (0.015)	20 (0.007)
2980.958	(16.241, 47.392)	(16.268, 47.381)	(16.304, 47.619)	(16.196, 48.132)	(16.202, 48.08)	(16.211, 48.124)	(16.206, 48.47)	(16.214, 48.083)	(16.251, 48.359)	(16.251, 48.362)
1585.129	(14.201, 47.442)	(14.22, 47.336)	(14.259, 47.534)	(14.135, 48.18)	(14.129, 47.388)	(14.139, 47.278)	(14.133, 47.424)	(14.139, 47.375)	(14.176, 47.953)	(14.176, 47.936)
842.895	(13.379, 49.057)	(13.4, 49.067)	(13.464, 49.183)	(13.317, 49.981)	(13.305, 49.313)	(13.313, 49.209)	(13.31, 49.466)	(13.312, 49.396)	(13.348, 49.893)	(13.349, 49.881)
448.211	(13.283, 49.489)	(13.306, 49.489)	(13.371, 49.691)	(13.218, 50.605)	(13.192, 49.802)	(13.2, 49.708)	(13.195, 49.853)	(13.189, 49.82)	(13.223, 50.437)	(13.224, 50.426)
238.337	(12.675, 48.203)	(12.691, 48.246)	(12.705, 48.394)	(12.634, 49.274)	(12.617, 48.925)	(12.625, 48.857)	(12.622, 48.992)	(12.618, 48.9)	(12.648, 49.573)	(12.649, 49.561)
126.736	(12.791, 51.678)	(12.804, 51.85)	(12.821, 51.639)	(12.757, 52.187)	(12.756, 53.047)	(12.765, 53.082)	(12.759, 53.512)	(12.753, 53.23)	(12.782, 53.336)	(12.782, 53.344)
67.392	(12.99, 47.300)	**(13.009, 47.213)**	(13.047, 47.37)	(12.941, 47.901)	(12.984, 47.415)	(12.991, 47.361)	(12.986, 47.613)	(12.994, 47.469)	(13.021, 47.791)	(13.02, 47.794)
35.836	(13.561, 52.193)	(13.559, 51.975)	(13.515, 51.59)	(13.477, 51.16)	(13.534, 51.194)	(13.541, 51.249)	(13.541, 51.16)	(13.558, 51.292)	(13.558, 50.83)	(13.588, 50.83)
19.056	(12.029, 68.203)	(12.035, 68.258)	(12.035, 67.988)	(11.94, 66.67)	(12.021, 67.544)	(12.028, 67.56)	(12.015, 68.187)	(12.028, 68.354)	(12.06, 68.077)	(12.058, 68.107)
10.133	(14.224, 81.347)	(14.235, 81.301)	(14.266, 81.324)	(14.184, 80.782)	(14.259, 80.037)	(14.267, 80.021)	(14.262, 80.026)	(14.272, 80.354)	(14.305, 80.168)	(14.302, 80.212)
5.388	(12.872, 54.095)	(12.895, 54.104)	(12.926, 53.702)	(12.87, 52.07)	(12.914, 53.514)	(12.919, 53.514)	(12.913, 53.945)	(12.92, 54.03)	(12.95, 53.677)	(12.949, 53.71)
2.865	(11.608, 79.496)	(11.64, 79.382)	(11.684, 79.465)	(11.583, 78.679)	(11.629, 78.551)	(11.634, 78.551)	(11.614, 78.531)	(11.618, 78.872)	(11.65, 78.364)	(11.647, 78.405)
1.524	(10.086, 75.889)	(10.114, 75.651)	(10.184, 75.592)	(10.051, 74.761)	(10.09, 74.363)	(10.095, 74.311)	(10.073, 74.413)	(10.075, 74.792)	(10.106, 74.1)	(10.103, 74.144)
0.81	(10.032, 79.517)	(10.051, 79.263)	(10.094, 79.325)	(9.983, 78.107)	(10.028, 78.435)	(10.033, 78.4)	(10.014, 78.353)	(10.019, 78.587)	(10.053, 78.176)	(10.049, 78.218)
0.431	(9.811, 78.784)	(9.837, 78.52)	(9.891, 78.6)	(9.79, 77.35)	(9.837, 77.661)	(9.843, 77.63)	(9.825, 77.613)	(9.818, 77.864)	(9.846, 77.454)	(9.846, 77.496)
0.229	(10.106, 83.871)	(10.115, 83.644)	(10.157, 83.668)	(10.042, 82.409)	(10.084, 82.672)	(10.09, 82.666)	(10.078, 82.637)	(10.073, 82.794)	(10.105, 82.672)	(10.102, 82.716)
0.122	(9.547, 86.89)	(9.548, 86.636)	(9.6, 86.675)	(9.535, 85.576)	(9.58, 85.272)	(9.589, 85.272)	(9.573, 85.225)	(9.549, 85.513)	(9.578, 85.992)	(9.576, 86.036)
0.065	(10.219, 86.494)	(10.223, 86.24)	(10.287, 86.264)	(10.184, 85.157)	(10.23, 84.971)	(10.239, 84.966)	(10.223, 84.91)	(10.207, 85.175)	(10.235, 85.479)	(10.233, 85.523)
0.034	(10.267, 85.835)	(10.268, 85.56)	(10.334, 85.58)	(10.242, 84.484)	(10.285, 84.148)	(10.294, 84.143)	(10.278, 84.112)	(10.262, 84.407)	(10.293, 84.797)	(10.291, 84.842)
0.018	(9.741, 84.284)	(9.741, 83.937)	(9.775, 83.957)	(9.733, 82.972)	(9.773, 82.141)	(9.781, 82.143)	(9.769, 82.121)	(9.756, 82.356)	(9.772, 83.187)	(9.77, 83.231)

References

D Adenyo, JR Guertin, B Candas, C Sirois, and D Talbot. Evaluation and comparison of covariate balance metrics in studies with time-dependent confounding, 2024a. URL https://arxiv.org/abs/2403.08577.

David Adenyo, Mireille E Schnitzer, David Berger, Jason R Guertin, Bernard Candas, and Denis Talbot. Longitudinal efficient adjustment sets for time-varying treatment effect estimation in nonparametric causal graphical models. *arXiv preprint arXiv:2410.01000*, 2024b.

H Bang and JM Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. ISSN 0006-341X.

P Bech, KB Stage, NPV Nair, JK Larsen, P Kragh-Sørensen, A Gjerris, Danish University Antidepressant Group, et al. The major depression rating scale (mds). inter-rater reliability and validity across different settings in randomized moclobemide trials. *Journal of Affective Disorders*, 42(1): 39–48, 1997.

Per Bech, N Timmerby, Klaus Martiny, M Lunde, and S Soendergaard. Psychometric evaluation of the major depression inventory (mdi) as depression severity scale using the lead (longitudinal expert assessment of all data) as index of validity. *BMC Psychiatry*, 15(1):1–7, 2015.

MA Brookhart, S Schneeweiss, KJ Rothman, RJ Glynn, J Avorn, and T Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156, 2006.

Michael Chaiton, Gisèle Contreras, Jennifer Brunet, Catherine M Sabiston, Erin O'Loughlin, Nancy CP Low, Igor Karp, Tracie A Barnett, and Jennifer O'Loughlin. Heterogeneity of depressive symptom trajectories through adolescence: Predicting outcomes in young adulthood. *J Can Acad Child Adolesc Psychiatry*, 22(2):96–105, 2013.

C Robert Cloninger. A systematic method for clinical description and classification of personality variants: A proposal. *Archives of general psychiatry*, 44(6):573–588, 1987.

SR Cole and MA Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–64, 2008.

Luis G Escobedo, Darrell G Kirch, and Robert F Anda. Depression and smoking initiation among us latinos. *Addiction*, 91(1):113–119, 1996.

Sybil BG Eysenck and Hans J Eysenck. Impulsiveness and venturesomeness: Their position in a dimensional system of personality description. *Psychological reports*, 43(3_suppl):1247–1255, 1978.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Susan Gruber and Mark J van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6 (1), 2010.

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

Holger Höfling, Harald Binder, and Martin Schumacher. A coordinate-wise optimization algorithm for the fused lasso. *arXiv preprint arXiv:1011.6409*, 2010.

JW Jackson. Diagnostics for confounding of time-varying and other joint exposures. *Epidemiology*, 27(6):859–869, 2016. doi: doi:10.1097/EDE.0000000000000547.

C Ju, S Gruber, SD Lendle, A Chambaz, JM Franklin, R Wyss, S Schneeweiss, and MJ van der Laan. Scalable collaborative targeted learning for high-dimensional data. *Statistical Methods in Medical Research*, 28(2):532–554, 2019. doi: 10.1177/0962280217729845.

C Ju, D Benkeser, and MJ van der Laan. Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*, 76(1):109–118, 2020. doi: 10.1111/biom.13121.

H Leeb and BM Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory, 21, 2005, 21-59*, 21:21–59, 2005.

G Lefebvre, JAC Delaney, and RW Platt. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine*, 27:3629–3642, 2008. doi: 10.1002/sim.3200.

Yan Liu, Mireille Schnitzer, Ronald Herrera, Iván Díaz, Jennifer O'Loughlin, and Marie-Pierre Sylvestre. The application of target trials with longitudinal targeted maximum likelihood estimation to assess the effect of alcohol consumption in adolescence on depressive symptoms in adulthood. *American journal of epidemiology*, July, 2023. Accepted.

WW Loh and S Vansteelandt. Confounder selection strategies targeting stable treatment effect estimators. *Statistics in Medicine*, 40(3):607–630, 2021.

R Neugebauer and M van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007. doi: https://doi.org/10.1016/j.jspi.2005.12.008. URL https://www.sciencedirect.com/science/article/pii/S0378375806000334.

Catherine Otter, Jörg Huber, and Adrian Bonner. Cloninger's tridimensional personality questionnaire: reliability in an english sample. *Personality and Individual Differences*, 18(4):471–480, 1995.

Jennifer O'Loughlin, Erika N Dugas, Jennifer Brunet, Joseph DiFranza, James C Engert, Andre Gervais, Katherine Gray-Donald, Igor Karp, Nancy C Low, Catherine Sabiston, et al. Cohort profile: the nicotine dependence in teens (ndit) study. *Int J Epidemiol*, 44(5):1537–1546, 2015.

E Persson, J Häggström, I Waernbaum, and X de Luna. Data-driven algorithms for dimension reduction in causal inference. *Computational Statistics & Data Analysis*, 105:280–292, 2017. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2016.08.012. URL https://www.sciencedirect.com/science/article/pii/S0167947316302018.

M Petersen, J Schwab, S Gruber, N Blaser, M Schomaker, and MJ Van der Laan. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2):147–185, 2014. ISSN 2193-3685.

Simon Racicot, Jennifer J McGrath, Igor Karp, and Jennifer O'Loughlin. Predictors of nicotine dependence symptoms among never-smoking adolescents: a longitudinal analysis from the nicotine dependence in teens study. *Drug and alcohol dependence*, 130(1-3):38–44, 2013.

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

M Rosenberg. Rosenberg self-esteem scale (rse). acceptance and commitment therapy. measures package, 61 (52), 18. *Wollongong, Australia: University of Wollongong*, 1965.

A Rotnitzky and E Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020.

A Rotnitzky, L Li, and X Li. A note on overadjustment in inverse probability weighted estimation. *Biometrika*, 97(4):997–1001, 2010.

JE Rudolph, D Benkeser, EH Kennedy, EF Schisterman, and AI Naimi. Estimation of the average causal effect in longitudinal data with time-varying exposures: The challenge of nonpositivity and the impact of model flexibility. *American Journal of Epidemiology*, 191(11):1962–1969, 2022. doi: 10.1093/aje/kwac136.

EF Schisterman, SR Cole, and RW Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488–495, 2009.

S Schneeweiss, JA Rassen, RJ Glynn, J Avorn, H Mogun, and MA Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522, 2009.

ME Schnitzer, MJ van der Laan, EE Moodie, and RW Platt. Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for clustered longitudinal data. *Annals of Applied Statistics*, 8(2):703–725, 2014. doi: 10.1214/14-aoas727.

ME Schnitzer, J Sango, S Ferreira Guerra, and MJ van der Laan. Data-adaptive longitudinal model selection in causal inference with collaborative targeted minimum loss-based estimation. *Biometrics*, pages 145–157, 2020. doi: 10.1111/biom.13135.

M Schomaker, MA Luque-Fernandez, V Leroy, and MA Davies. Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Statistics in Medicine*, 38(24):4888–4911, 2019.

SM Shortreed and A Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017. doi: doi:10.1111/biom.12679.

D Talbot, G Lefebvre, and J Atherton. The bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2):207–236, 2015.

D Tang, D Kong, W Pan, and L Wang. Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics*, page 1–12, 2022. doi: 10.1111/biom.13625.

R Tibshirani, M Saunders, S Rosset, J Zhu, and K Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, 2005.

Ryan J Tibshirani. *The solution path of the generalized lasso*. Stanford University, 2011.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

MJ Van der Laan and S Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1), 2012.

MJ van der Laan and S Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *International Journal of Biostatistics*, 8(1):10.1515/1557–4679.1370, 2012.

V Viallon, S Lambert-Lacroix, H Höfling, and F Picard. Adaptive generalized fused- lasso: Asymptotic properties and applications. pages hal–00813281, 2013.

V Viallon, S Lambert-Lacroix, H Hoefling, and F Picard. On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, 26:285–301, 2016. doi: https://doi.org/10.1007/s11222-014-9497-6.

C Wang, G Parmigiani, and F Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–671, 2012.

Thomas Ashby Wills, Michael Windle, and Sean D Cleary. Temperament and novelty seeking in adolescent substance use: convergence of dimensions of temperament with constructs from cloninger's theory. *Journal of Personality and Social Psychology*, 74(2):387, 1998.

A Wilson and BJ Reich. Confounder selection via penalized credible regions. *Biometrics*, 70: 852–861, 2014.

H Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.