Wenjing Zheng^{1,2} / Zhehui Luo³ / Mark J van der Laan^{1,2}

Marginal Structural Models with Counterfactual Effect Modifiers

¹ Division of Biostatistics, University of California, Berkeley, USA² Center for Targeted Learning, University of California, Berkeley, USA³ Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, USA, E-mail: zluo@epi.msu.edu

Abstract:

In health and social sciences, research questions often involve systematic assessment of the modification of treatment causal effect by patient characteristics. In longitudinal settings, time-varying or post-intervention effect modifiers are also of interest. In this work, we investigate the robust and efficient estimation of the Counterfactual-History-Adjusted Marginal Structural Model (van der Laan MJ, Petersen M. Statistical learning of origin-specific statically optimal individualized treatment rules. *Int J Biostat.* 2007;3), which models the conditional intervention-specific mean outcome given a counterfactual modifier history in an ideal experiment. We establish the semiparametric efficiency theory for these models, and present a substitution-based, semiparametric efficient and doubly robust estimator using the targeted maximum likelihood estimation methodology (TMLE, e.g. van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. *Int J Biostat.* 2006;2, van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data, 1st ed. Springer Series in Statistics. Springer, 2011). To facilitate implementation in applications where the effect modifier is high dimensional, our third contribution is a projected influence function (and the corresponding *projected* TMLE estimator), which retains most of the robustness of its efficient peer and can be easily implemented in applications where the use of the efficient influence function becomes taxing. We compare the *projected* TMLE estimator with an Inverse Probability of Treatment Weighted estimator (e.g. Robins JM. Marginal structural models. In: Proceedings of the American Statistical Association. Section on Bayesian Statistical Science, 1-10. 1997a, Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11:561–570), and a non-targeted G-computation estimator (Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Math Modell.* 1986;7:1393–1512.). The comparative performance of these estimators is assessed in a simulation study. The use of the *projected* TMLE estimator is illustrated in a secondary data analysis for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial where effect modifiers are subject to missing at random.

Keywords: causal inference, effect modification, epidemiology, machine learning**DOI:** 10.1515/ijb-2018-0039**Received:** April 3, 2018; **Accepted:** May 9, 2018

1 Introduction

In social and health sciences, research questions often involve systematic comparison of the effectiveness of different longitudinal exposures or treatment strategies on an outcome of interest. Specifically, consider a study where individuals are followed over time. In addition to their baseline covariates, we record their time-varying treatments, time-varying covariates, and the outcomes of interest. Time-varying confounding is ubiquitous in these settings: the treatment of interest depends on past covariates and in turn affects future covariates. Right censoring is often present, in response to past covariates and treatments. It has been widely recognized that in these cases, conventional analytic methods, such as multiple regressions, often fail to properly account for the time-varying confounding of the treatment effect (e.g. [7]). Marginal Structural Models (MSMs), introduced by [4], are well-established and widely used tools to address the problem of time-varying confounding. These models estimate the marginal expectation of an *intervention-specific counterfactual outcome*, i.e. the mean outcome of a subject in an ideal experiment where he/she was assigned to follow a given intervention.

To assess effect modification, MSMs are traditionally used to model the conditional counterfactual mean outcome given an observed history. Yet, in many settings one may wish to model the conditional counterfactual mean outcome given a counterfactual history. Consider the simple case of effect modification by base-

Zhehui Luo is the corresponding author.

© 2018 Walter de Gruyter GmbH, Berlin/Boston.

line covariates. A traditional *observed baseline MSM* conditions on the observed baseline modifiers [7] and allows one to assess how the treatment effect changes as a function of the observed modifier values. (*Observed*) *History-Adjusted MSMs* (HA-MSMs), introduced in [8, 9], generalize the *observed baseline MSMs* by modeling the counterfactual mean outcome given the observed history of treatment and modifiers of interest up till a time point.

However, since the modifiers of interest may be affected by their preceding treatment assignments, which may in turn depend on past modifiers and other covariates, the counterfactual mean within each stratum of this history will also be affected by the *observed* treatment mechanism. In this case, the parameters of the HA-MSM would not be generalizable to an equivalent population with different or *counterfactual* treatment mechanisms [10]. We consider a model in which the true outcome that one wishes to model is the conditional mean outcome given modifier history in an ideal experiment where subjects were assigned a given intervention of interest and a *counterfactual* history of the time-varying modifiers of interest.

To model these conditional counterfactual mean outcomes given a *counterfactual* history, we can use the so-called Counterfactual History Adjusted MSMs (CHA-MSM), introduced by [1]. Inverse Probability of Treatment Weighted (IPTW) estimators for CHA-MSM were proposed in [1]. These estimators are very intuitive, can be easily implemented using standard software, and offer influence function based standard error estimates. However, their consistency relies on consistent estimation of all the treatment weights. Doubly Robust IPTW (DR-IPTW) estimators for CHA-MSM were also described in [1]; they were based on estimating equations derived by orthogonalizing the IPTW estimating function with respect to the treatment mechanism. These DR-IPTW estimators are consistent if the treatment weights or the conditional covariate and outcome densities are consistently estimated. Moreover, they are solutions to the estimating equation defined by the efficient influence function (EIF), and thus are asymptotically semi-parametric efficient.

Despite these advances, there are still many gaps in the efficiency theory and robust estimation of CHA-MSMs. Firstly, even with the EIF being a key actor in semi-parametric estimation, there still lacks an explicit representation of it as an orthogonal decomposition of the nuisance parameters corresponding to the time-varying covariates. Compared to the IPTW-orthogonalized representation in [1], such an explicit representation would provide a comprehensive picture of the efficiency theory for CHA-MSM. In particular, by shining a light on the role of the nuisance parameters in the EIF, such an explicit representation can inform the study of semi-parametric estimation for these models, advise on the trade-offs in estimating different nuisance parameters, and provide insights on the challenges and solutions to handling high-dimensional covariates. Secondly, estimating equation based estimators, e.g. the IPTW and the DR-IPTW, may be unstable in the presence of near positivity violations [11], resulting in biased point and standard error estimates. In applications with dynamic treatment regimes, this instability is especially difficult to circumvent due to the limitations of effective weight stabilization. In contrast, a substitution-based estimator for these models can provide a way to maximize finite sample performance by preserving global information about the parameters.

This paper aims to fill these gaps in the literature by establishing the efficiency theory for CHA-MSM and providing substitution-based, semi-parametric efficient and robust estimators. Firstly, we describe the identification of the conditional counterfactual intervention-specific mean outcome given a counterfactual history, and the identification of the corresponding MSM parameters of interest. Secondly, we determine the EIF for these statistical parameters as an orthogonal decomposition of the nuisance parameters. This EIF is used to construct a substitution-based, semi-parametric efficient and doubly robust estimator using the targeted maximum likelihood estimation methodology (TMLE, e.g. [2, 3]). However, as we shall see, due to the form of the EIF, the computations in this estimator may prove arduous in applications where the effect modifiers are high dimensional. To address this problem, our third contribution is a *projected* influence function (and the corresponding TMLE estimator), which retains most of the robustness of its efficient peer and can be used in applications where the use of the EIF becomes taxing.

1.1 Illustrative example

Throughout this paper, we will illustrate our presentation using an example from mental health research. The STAR*D (Sequenced Treatment Alternatives to Relieve Depression) trial is a multi-level, longitudinal pragmatic trial of treatment strategies for major depression. After an initial screening, patients are enrolled into level 1 of the study, where everyone is assigned citalopram. At the start of each subsequent level, if a patient still remains in the study, then he/she is randomized to an option within one of the chosen treatment strategies. At level 2, the patient and his/her provider can choose between augmenting the citalopram with multiple options or switching to a new regimen with multiple options. We are interested in the comparative effect of augmenting vs switching medication. Because these two strategies are not randomized, this analysis is analogous to an observational study.

Suppose we wish to assess the effect modification of level 2 treatment strategy (augmenting vs switching) by the depression symptoms measured prior to entering level 2. Covariates either measured at enrollment or level 1 exit, for our purpose, are both considered baseline modifiers. These symptom measures are obtained at clinical visits and the level 1 exit survey. It is reasonable to believe that the more depressed patients and those less satisfied with their level 1 treatment may be less inclined to follow up with clinical visits or to respond to surveys. In this case, a simple complete-case analysis of effect modification may only provide results applicable to patients with less severe symptoms or more satisfied with their level 1 treatment. If we assume that this missingness, while associated with the outcome, can be predicted using covariate history collected up to level 1 exit, then we can regain some of the generalizability. Indeed, we cast these symptom measures as counterfactual variables under an intervention on their missingness mechanism. This way, the target parameter is formulated in terms of an ideal experiment where the report of the symptom measures was always ensured (i.e., intervened to be non-missing).

1.2 Organization of the article

This paper is organized as follows. In Section 2, we formulate the causal inference problem and determine the identifiability of the desired causal parameters from the observed data distribution. In Section 3, we present the EIF for the parameter of interest under a saturated semiparametric model, as well as a projected influence function. The robustness conditions for the efficient and projected influence functions are also established. We then present the construction of two TMLE estimators, one using the EIF (we call it the *full* TMLE) and one using the projected influence function (we call it the *projected* TMLE). For comparison we presented an IPTW estimator and a non-targeted substitution estimator. In Section 4, a simulation study demonstrates robustness properties of the *projected* TMLE. In Section 5, we use the STAR*D trial to illustrate the application of the *projected* TMLE. A final summary concludes the article.

2 Data structure and parameters of interest

For simplicity of exposition, we consider the case where one wishes to assess the conditional intervention-specific mean outcome given the counterfactual history *up to the first time point*.

2.1 Data structure

Specifically, consider a longitudinal data structure

$$O = (W, A_0, V, L_0, A_1, L_1, \dots, A_K, L_K) \sim P_0,$$

where W encodes baseline covariates, A_t encodes the exposures of interest and censoring indicators at time t , V encodes the time-varying history (between treatment A_0 and A_1) on which one wishes to condition, L_t encodes covariates (including time-varying confounders and the outcome process Y_t) measured between A_t and A_{t+1} . Variable Y_K is the final outcome of interest. For the sake of discussion, assume that Y_K is either a binary or a bounded continuous variable (without losing generality, we may assume it is bounded in $(0,1)$). We note that in general, A_t encodes intervention variables of interest, i.e. variables whose distribution one would set in an ideal experiment. These include exposure variables (where one would randomize in an ideal experiment) as well as censoring variables (where one would prevent in an ideal experiment).

For later convenience, we introduce some useful notations. For a time-varying variable X_s , let $\mathbf{X}_s \equiv (X_1, \dots, X_s)$ and $\mathbf{X}_{s,t} \equiv (X_s, \dots, X_t)$ for $s < t$. We will also use the shorthand $\mathbf{X} = \mathbf{X}_K$. The equivalent notations in lower case will apply to the realizations of the corresponding variable. Finally, variables with degenerate indices, such as L_t at $t = -1$, are empty.

The time-ordering assumptions can be captured by a nonparametric structural equations model (NPSEM, [12]):

$$\begin{aligned} W &= f_W(U_W); \quad A_0 = f_{A_0}(W, U_{A_0}); \quad V = f_V(W, A_0, U_V); \quad L_0 = f_{L_0}(W, A_0, V, U_{L_0}); \\ A_t &= f_{A_t}(W, A_0, V, \mathbf{A}_{t-1}, \mathbf{L}_{t-1}, U_{A_t}); \quad L_t = f_{L_t}(W, A_0, V, \mathbf{A}_t, \mathbf{L}_{t-1}, U_{L_t}). \end{aligned} \quad (1)$$

This framework assumes that each variable X in the observed data structure is an unknown deterministic function of all its preceding observed variables and some unmeasured exogenous random factors U . We will refer to the observed variables in the input of f_X as the parents of X and denote their set as $Pa(X)$.

Concretely, we describe the data example introduced in Section 1.1 with these notations. Our goal is to assess the effect modification of level 2 treatment strategy (augmenting vs switching) by the depression symptoms measured prior to entering level 2; yet many of these baseline effect modifiers were missing at random. By study protocol, all subjects will have either entered remission (treatment success), moved onto the next level (treatment failure), or dropped out (right censoring), by the end of 23 weeks of the level 2 period. For our study goal, the variables that we would enforce/randomize in an ideal experiment are the missingness status of the baseline effect modifier, the treatment assignment at the start of level 2, and the censoring status in each of the 23 weeks.

To this end, let V be the baseline effect modifier of interest. Let A_0 be the indicator for non-missingness of V , and A_1 be the treatment strategy received by a patient at level 2 ($A_1 = 0$ for augmenting medication and $A_1 = 1$ for switching medication). We use W to encode the baseline covariates that may affect the measurement status A_0 , and L_0 to encode other baseline covariates (beside the modifier of interest) measured prior to treatment assignment A_1 . Hence, under this indexing, $K = 24$ and the first week of level 2 starts at $t = 2$. For $t \geq 2$, A_t is a counting process which drops to 0 if patient was censored by time t , and L_t are time-varying covariates such as visit statistics (duration in level thus far, visit frequency, etc), side-effect burden and symptom measures at time t . The variable L_t also contains two counting processes: the outcome process Y_t , which is a binary indicator for entering remission by time t , and an exit process E_t that jumps to 1 if a patient is moved to the next level, in which case the remission status will be zero for this level and the patient is considered non-censored (since the outcome was observed to be unsuccessful). Our final outcome of interest is Y_K — the remission status by the end of 23 weeks.

2.2 Parameter of interest

To assess the effect of the interventions, we study the *intervention-specific counterfactual mean outcomes* as a function of the interventions. These counterfactual mean outcomes can be obtained from an ideal experiment where one intervenes to assign $\mathbf{A} = \mathbf{a}$, and measures the resulting covariates and the outcome of interest. In our example, an ideal experiment would set $A_0 = 1$ (always measure baseline modifier), $A_1 = a_1$ (switching or augmenting), and $\mathbf{A}_{2,K} = 1$ (always prevent drop outs). This data-generating process can be formalized using eq. (1) by setting $A_t = a_t$ in the equations for A_t and in the parents of the non-intervention variables V , L_t and Y . We denote these resulting, possibly *counterfactual*, covariates as $V(a_0)$, $L_t(\mathbf{a}_t)$ and $Y_K(\mathbf{a})$.

Now, we wish to assess effect modification of such treatment effects by changing values of V . That is, we ask “how does the differential effect of $(a_0, \mathbf{a}_{1,K})$ vs $(a_0, \mathbf{a}_{1,K})$, and of $(a_0, \mathbf{a}_{1,K})$ vs $(a_0^*, \mathbf{a}_{1,K})$, differ as a function of V , which is affected by the treatment assignment A_0 ?”. To answer this question, our parameters of interest are the so-called *Counterfactual-History-Adjusted* mean outcomes [1]

$$\rho(\mathbf{a}, v) \equiv E(Y(\mathbf{a}) \mid V(a_0) = v), \quad (2)$$

and our goal is to study how these mean outcomes change as a function of a_0 , $\mathbf{a}_{1,K}$ and v . As discussed in Section 1, eq. (2) differs from the traditional History-Adjusted mean outcomes, $E(Y(A_0, \mathbf{a}_{1,K}) \mid V = v, A_0 = a_0)$, in that these traditional parameters condition on the observed treatment and modifier values. Hence within each such stratum, the conditional mean outcome may still depend on the observed treatment mechanism. As such, these parameters are affected by potential selection bias.

In our example, the conditional mean outcomes we wish to study are $E(Y(1, \mathbf{a}_{1,K}) \mid V(1) = v)$, with $\{\mathbf{a}_{1,K} = (a_1, \dots, a_K) : a_1 \in \{0, 1\}; a_t = 1, t \geq 2\}$. In words, it is the *intervention-specific counterfactual mean outcome* in an ideal experiment without missing measurements or censoring.

A challenge to study of eq. (2) is that the curse of dimensionality arises in applications with more than two time points and/or with categorical or continuous V . To address this issue, it is useful to summarize eq. (2) using a working MSM, $m_\psi(\mathbf{a}, v)$, which is parametrized by a finite dimensional $\psi \in S \subset \mathbb{R}^d$. Since the outcome Y is binary (or bounded within the unit interval), concretely we consider a logistic MSM

$$m_\psi(\mathbf{a}, v) = \text{expit}(\psi \cdot \phi(\mathbf{a}, v)), \quad (3)$$

where $\phi(\mathbf{a}, v)$ is the vector of linear predictors in the generalized linear model. The methods presented here are easily modified to other forms of MSM.

Given a CHA-MSM eq. (3) for the conditional mean outcomes eq. (2), the true MSM parameter ψ can be interpreted as the best summary measure of $\rho(\mathbf{a}, v)$ as a function of $\mathbf{a}$ and v . Formally, let $\mathcal{A}$ denote the set of interventions of interest and $\mathcal{V}$ denote the set of modifier values of interest. For a given kernel weight function

$h(\mathbf{a}, v)$, the true MSM parameter ψ in eq. (3) is defined as

$$\psi = \arg \min_{\psi \in S} \left\{ - \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} P(V(a_0) = v) h(\mathbf{a}, v) \right. \\ \left. \times \left\{ \rho(\mathbf{a}, v) \log m_\psi(\mathbf{a}, v) + (1 - \rho(\mathbf{a}, v)) \log (1 - m_\psi(\mathbf{a}, v)) \right\} \right\}. \quad (4)$$

In words, ψ yields the best weighted approximation of the counterfactual conditional dose-response curve $\rho(\mathbf{a}, v)$, according to the quasi-loglikelihood loss, kernel weights and working model $m_\psi(\mathbf{a}, v)$.

The rest of this paper is devoted to the identification and inference of eq. (4) from the observed data.

2.3 Statistical estimand

To identify eq. (4) from the data generating distribution P_0 , we make the Positivity Assumption (PA) and the Sequential Randomization Assumption (SRA, derived by [13]). Specifically, under the PA, there exists $\alpha_t > 0$ such that $\alpha_t \leq P_0(A_t = a_t \mid Pa(A_t))$, for all t and $\mathbf{a} \in \mathcal{A}$, almost everywhere. The SRA assumes that $A_t \perp (V(a_0), \{L_t(\mathbf{a}) : t\}) \mid Pa(A_t)$. Under these conditions, the joint distribution $(V(a_0), \{L_t(\mathbf{a}) : t\})$ is identifiable from the observed data distribution P_0 . In our STAR*D example, the plausibility of the SRA can be fortified by measuring sufficiently many confounders of the modifier's missingness, the treatment selection, and the censoring mechanism.

By calculations shown in the Appendix, the SRA allows us to identify $P(V(a_0) = v)$ as

$$\gamma(a_0, v) \equiv E_W \{P(V = v \mid A_0 = a_0, W)\}, \quad (5)$$

and the counterfactual mean outcome $\rho(\mathbf{a}, v)$ as

$$\rho(\mathbf{a}, v) \equiv E_W \left\{ \frac{P(V = v \mid A_0 = a_0, W)}{\gamma(a_0, v)} \times Q_0^{\mathbf{a}}(V = v, W) \right\}, \quad (6)$$

where, for $t = 0, \dots, K$,

$$Q_t^{\mathbf{a}}(\mathbf{L}_{t-1}, V, W) \equiv \sum_{\mathbf{l}_{t,K}} y_K \left(\prod_{j=t}^K P(l_j \mid \mathbf{A}_j = \mathbf{a}_j, \mathbf{L}_{t-1}, \mathbf{l}_{t,j-1}, V, A_0 = a_0, W) \right), \quad (7)$$

where $\mathbf{l}_{t,j-1} = (l_t, \dots, l_{j-1})$ for $t < j-1$. Equation (7) represents a conditional expectation of the final outcome Y_K given observed histories $(\mathbf{L}_{t-1}, V, W)$ and under an intervention $\mathbf{a}$ on $\mathbf{A}$.

For simplicity, we suppressed the notation for the functionals $\gamma, \rho, Q_t^{\mathbf{a}}$, as each should be evaluated with respect to a given distribution. We will use $\gamma(P), \rho(P), Q_t^{\mathbf{a}}(P)$ to express explicitly these functionals evaluated at a specific distribution P when needed, but will keep the shorthand notation otherwise.

In comparison to eq. (6), parameters in conventional settings where the modifier V is not counterfactual can be written as

$$E_W \left\{ \frac{1/P(A_0 = a_0 \mid W)}{E_W(1/P(A_0 = a_0 \mid W) \mid V = v, A_0 = a_0)} Q_1^{\mathbf{a}}(L_0, V = v, W) \mid V = v, A_0 = a_0 \right\}.$$

The weight $\frac{1/P(A_0 = a_0 \mid W)}{E_W(1/P(A_0 = a_0 \mid W) \mid V = v, A_0 = a_0)}$ adjusts for potential selection bias introduced by treatment assignment $A_0 = a_0$. Indeed, if A_0 does not depend on W (in our STAR*D example, this means missingness is completely at random), then this weight equals 1, in which case $\rho(\mathbf{a}, v)$ is equivalent to the estimand in an analysis stratified by A_0 .

Combining eqs. (5) and (6), the causal MSM parameter ψ in eq. (4) identifies to

$$\psi_0 \equiv \Psi(P_0) \equiv \arg \min_{\psi \in S} \left\{ - \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \gamma(a_0, v) h(\mathbf{a}, v) \right. \\ \left. \times \left\{ \rho(\mathbf{a}, v) \log m_\psi(\mathbf{a}, v) + (1 - \rho(\mathbf{a}, v)) \log (1 - m_\psi(\mathbf{a}, v)) \right\} \right\}, \quad (8)$$

where the functionals γ, ρ are evaluated at P_0 . In the forthcoming sections, we study the statistical inference of $\Psi(P_0)$.

2.4 Notations

Before we proceed, let us introduce some more useful definitions and notations. Let $\mathcal{M}$ be a saturated semi-parametric model containing our data generating distribution P_0 . The parameter of interest in eq. (8) is the map $P \mapsto \Psi(P)$, from $\mathcal{M}$ to $\mathbb{R}^d$, evaluated at P_0 .

Suppose we observe n i.i.d. copies of $O \sim P_0$. Let P_n denote the empirical distribution of this sample. For a function f of O , we will write $P_n f \equiv \frac{1}{n} \sum_{i=1}^n f$, and for a distribution P , we will write $Pf = E_P f(O)$.

Bang and Robins [14] noted the following recursive property when dealing with longitudinal intervention-specific mean outcomes:

$$Q_t^a(\mathbf{L}_{t-1}, V, W) = E_P [Q_{t+1}^a(\mathbf{L}_t, V, W) | \mathbf{A}_t = \mathbf{a}_t, \mathbf{L}_{t-1}, V, A_0 = a_0, W], \quad (9)$$

for $t = 0, \dots, K$. This will prove useful in our upcoming endeavor. We also adopt the notations $Q^W(P)$ for the marginal distribution of W , $Q^V(P)$ for the conditional distribution $P(V | W, A_0)$, and $Q(P) \equiv (Q^W(P), Q^V(P), Q^a(P) : t = 0, \dots, K)$. We write g for the treatment allocation probabilities $P(A_t | \mathbf{A}_{t-1}, \mathbf{L}_{t-1}, V, W)$. When referring to a generic $P \in \mathcal{M}$, we may sometimes write Q and g in place of $Q(P)$ and $g(P)$, similarly for their respective components. When referring to the functionals at the data-generating distribution P_0 , we may sometimes write Q_0 and g_0 in place of $Q(P_0)$ and $g(P_0)$. When there is no confusion, we will simply write Q and g under the implied distribution.

3 Statistical inference: A tale of two influence functions

As defined in eq. (8), $\Psi(P)$ optimizes a function of the functionals γ and ρ evaluated at P . Note also that $\rho(\mathbf{a}, v) = \frac{\eta(\mathbf{a}, v)}{\gamma(a_0, v)}$, where

$$\eta(\mathbf{a}, v) \equiv E \{P(V = v | A_0 = a_0, W) \times Q_0^a(V = v, W)\}.$$

We will make use of the following useful characterizations for $\Psi(P)$:

Remark 1

For $m_\psi(\mathbf{a}, v) = \text{expit}(\psi \cdot \phi(\mathbf{a}, v))$, $\Psi(P)$ defined in eq. (8), and η and γ functionals evaluated at P , we have

$$\begin{aligned} 0 &= U(\Psi(P)) \equiv \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) \gamma(a_0, v) \left(\frac{\eta(\mathbf{a}, v)}{\gamma(a_0, v)} - m_\Psi(\mathbf{a}, v) \right) \\ &= E_W \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \tilde{h}(\mathbf{a}, v) P(V = v | A_0 = a_0, W) (Q_0^a(V = v, W) - m_\Psi(\mathbf{a}, v)) \right\} \\ &= E_W \left\{ \sum_{\mathbf{a} \in \mathcal{A}} E_V [\tilde{h}(\mathbf{a}, V) (Q_0^a(V, W) - m_\Psi(\mathbf{a}, V)) | A_0 = a_0, W] \right\} \\ &= E \left\{ \sum_{\mathbf{a} \in \mathcal{A}} \frac{I(A_0 = a_0)}{g(A_0 = a_0 | W)} \tilde{h}(\mathbf{a}, V) \{Q_0^a(V, W) - m_\Psi(\mathbf{a}, V)\} \right\}, \end{aligned}$$

where $\tilde{h}(\mathbf{a}, v) \equiv h(\mathbf{a}, v) \phi(\mathbf{a}, v)$, and the last expectation is taken over (W, A_0, V) .

The derivations are straightforward, and we left them in the Appendix for reference.

Suppose that the following normalizing matrix is invertible at $\psi = \Psi$:

$$\begin{aligned} M(\psi) &= \frac{\partial}{\partial \psi} \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \gamma(a_0, v) h(\mathbf{a}, v) \phi(\mathbf{a}, v) \{\rho(\mathbf{a}, v) - m_\psi(\mathbf{a}, v)\} \\ &= E \left\{ \sum_{\mathbf{a} \in \mathcal{A}} \frac{I(A_0 = a_0)}{g(a_0 | W)} h(\mathbf{a}, V) \phi(\mathbf{a}, V) \phi(\mathbf{a}, V)^\top m_\psi(\mathbf{a}, V) [1 - m_\psi(\mathbf{a}, V)] \right\}. \end{aligned} \quad (10)$$

3.1 Inverse probability weighted estimator ψ^{IPW}

Define

$$C_t^{\mathbf{a}}(g) \equiv \frac{I(\mathbf{A}_t = \mathbf{a}_t)}{g(A_0 = a_0 | W) \prod_{j=1}^t g(A_j = a_j | \mathbf{A}_{j-1} = \mathbf{a}_{j-1}, \mathbf{L}_{j-1}, V, W)},$$

for $t = 0, \dots, K$. We note that these are functionals of $g(P)$, but we have suppressed the notation here.

From Remark 1, a valid estimating function for Ψ is given by

$$D^{IPW}(g, \cdot) \equiv \sum_{\mathbf{a} \in \mathcal{A}} C_K^{\mathbf{a}}(g) \tilde{h}(\mathbf{a}, V) (Y_K - m_{\psi}(\mathbf{a}, V)). \quad (11)$$

Up to a normalizing matrix $M(\psi)^{-1}$, $D^{IPW}(g, \psi)$ is a gradient for Ψ under a model $\mathcal{M}_g$ where g is known. Note that $P_0 D^{IPW}(g(P_0), \psi) = U(\psi_0)$ and the corresponding equation to zero is solved for $\psi_0 = \Psi(P_0)$. Therefore, this is an unbiased estimating function for ψ_0 and ψ^{IPW} is an unbiased estimator if $g = g(P_0)$.

To implement the IPW estimator, we first obtain estimators of the probability of treatment propensity $\hat{g}$ of g . For an observation in the dataset with $\mathbf{A}$ having a value in $\mathcal{A}$, the variable $C_K^{\mathbf{A}}(\hat{g})$ will have numerator equals 1 and probabilities in the denominator given by $\hat{g}(A_j | \mathbf{A}_{j-1}, \mathbf{L}_{j-1}, V, W)$; otherwise $C_K^{\mathbf{A}}(\hat{g}) = 0$. The estimator ψ^{IPW} can be obtained by fitting a weighted logistic regression of Y_K on $\phi(\mathbf{A}, V)$, with weights $C_K^{\mathbf{A}}(\hat{g}) \tilde{h}(\mathbf{A}, V)$. This ψ^{IPW} satisfies $P_n D^{IPW}(\hat{g}, \psi^{IPW}) = 0$, and it is consistent if $\hat{g}$ consistently estimates g . Under standard regularity and empirical process conditions, ψ^{IPW} is asymptotically linear with influence function $M(\psi_0)^{-1} D^{IPW}(g, \psi_0)$. The asymptotic covariance of $\sqrt{n}(\psi^{IPW} - \psi_0)$ can be estimated by the sample covariance matrix Σ^{IPW} of $\{M(\psi^{IPW})^{-1} D^{IPW}(\hat{g}, \psi^{IPW})\}$.

Due to its inverse weighting by products of treatment and censoring probabilities, this estimator may be sensitive to near positivity violations, wherein observations with small observed treatment propensity would be given excessive weight. This can be particularly pronounced as the number of time points increase. Substitution estimators like the ones we propose in sections 3.2 and 3.3 can slightly mitigate this problem by incorporating global information in the parameter map. However, the effect of near positivity violations can still take form of poor estimation in the nuisance parameters.

3.2 Non-targeted substitution estimators ψ_V^{subs} and $\psi_{A_0}^{subs}$

From eq. (8) and Remark 1, we can formulate Ψ as $\Psi(Q_0^{\mathbf{a}}, Q^V)$ or $\Psi(Q_0^{\mathbf{a}}, g)$. These two formulations lead to two non-targeted substitution estimators, ψ_V^{subs} and $\psi_{A_0}^{subs}$. Note the latter option opens the door for G-computation estimators in applications with high-dimensional V .

We start by obtaining an estimator of $Q_0^{\mathbf{a}}(V, W)$ for a given $\mathbf{a} \in \mathcal{A}$. We will exploit the recursive property of $Q_t^{\mathbf{a}}$ noted in eq. (9), so that we can avoid estimation of densities, and will instead employ regression-based procedures to estimate the conditional expectations.

1. Initiate at $t = K$: We first regress Y_K on $(\mathbf{L}_{K-1}, \mathbf{A}, V, W)$ among observations uncensored by time K , to obtain an estimator of $E(Y_K | \mathbf{L}_{K-1}, \mathbf{A}, V, W)$. This can be implemented using data-adaptive (e.g. SuperLearner by [15]) or parametric procedures. We then evaluate this fitted estimator at the observed $(\mathbf{L}_{K-1}, V, W)$ with setting $\mathbf{A} = \mathbf{a}$, to obtain an estimator $\hat{Q}_K^{\mathbf{a}}(\mathbf{L}_{K-1}, V, W)$ for each observation uncensored at time $K - 1$.
2. At each $t = K - 1, \dots, 0$, in decreasing order, we would have obtained an estimator $\hat{Q}_{t+1}^{\mathbf{a}}(\mathbf{L}_t, V, W)$ for observations uncensored by time $t + 1$. We now regress this variable on $(\mathbf{L}_{t-1}, \mathbf{A}_t, V, W)$ among observations uncensored by time t , and then evaluate the resulting fit at the observed $(\mathbf{L}_{t-1}, V, W)$ with setting $\mathbf{A}_t = \mathbf{a}_t$ among observations uncensored at time $t - 1$.
3. After running step 2 sequentially from $t = K - 1$ down to $t = 0$, we would have an estimator $\hat{Q}_0^{\mathbf{a}}(V, W)$ for each of the n observations. We can now use $\hat{Q}_0^{\mathbf{a}}(V, W)$ to estimate Ψ through the characterizing equations in Remark 1.

Consider the first representation $\Psi(Q_0^{\mathbf{a}}, Q^V)$. Let $\hat{Q}^V$ be an estimators of $Q^V(v | A_0 = a_0, W)$. For each $\mathbf{a} \in \mathcal{A}$ and each $v \in \mathcal{V}$, we create a row of data consisting of $\hat{Q}_0^{\mathbf{a}}(v, W)$, $\phi(\mathbf{a}, v)$, $h(\mathbf{a}, v)$, $\hat{Q}^V$ for each observation in the dataset. We then pool together these datasets characterized by $(\mathbf{a}, v)$. The estimator ψ_V^{subs} can be obtained by a weighted logistic regression of $\hat{Q}_0^{\mathbf{a}}(v, W)$ onto $\phi(\mathbf{a}, v)$ on the pooled dataset, with weights $h(\mathbf{a}, v) \hat{Q}^V$. This ψ_V^{subs} is consistent if both $\hat{Q}_0^{\mathbf{a}}$ and $\hat{Q}^V$ are consistent. Here we use the empirical distribution to estimate the marginal distribution of W .

Consider now the alternative representation $\Psi(Q_0^a, g_0)$, from the last equality in Remark 1. Let $\hat{g}$ be an estimator of g . For each $\mathbf{a} \in \mathcal{A}$, we create a row of data consisting of $\hat{Q}_0^a(V, W)$, $\phi(\mathbf{a}, V)$, $h(\mathbf{a}, V)$ and $I(A_0 = a_0)/\hat{g}(a_0 | W)$ for each observation in the dataset. We then pool together these datasets characterized by $\mathbf{a}$. The estimator $\psi_{A_0}^{subs}$ can be obtained by a weighted logistic regression of $\hat{Q}_0^a(V, W)$ onto $\phi(\mathbf{a}, V)$ on the pooled dataset, with weights $h(\mathbf{a}, V) \frac{I(A_0=a_0)}{\hat{g}(a_0|W)}$. This $\psi_{A_0}^{subs}$ is consistent if both $\hat{Q}_0^a$ and $\hat{g}$ are consistent. Here we use the empirical distribution to estimate the marginal distribution of (W, A_0, V) .

While these substitution estimators use weighted regressions to obtain estimators for ψ_0 , they are less susceptible to extremely large weights than the IPW estimator described above. Indeed, $\hat{Q}_0^a$ is bounded within $(0, 1)$ while the inverse $1/\hat{g}$ utilizes product of only one conditional probability. If the Q components are correctly specified parametric models, then the substitution estimators are also asymptotically linear and converge to a normal random variable at $n^{1/2}$. However, the parametric modeling assumptions are difficult to justify in scientific investigations. Data-adaptive alternatives may yield estimators which are not $n^{1/2}$ consistent, which prohibits statistical inference.

3.3 Targeted maximum likelihood estimators

In a non-targeted substitution estimator, we can estimate the nuisance parameters of the data-generating distribution by stratification (nonparametric MLE), by fitting a parametric statistical model, or by using a machine learning algorithm. These estimates are then used to evaluate the parameter of interest. As the number of potential confounders increases, these methods may break down due to the curse of dimensionality, or yield a bias-variance trade off that is not optimal for the parameter of interest. The targeted learning approach aims at developing optimal ($n^{1/2}$ consistent, asymptotically normal, and efficient) estimators of smooth low-dimensional parameters through the use of data-adaptive machine learning [16].

A targeted MLE adds a loss-based targeting step that aims to modify the initial estimators towards the parameter of interest, and provides potential robustness and semiparametric efficiency gains. The targeting is guided by an influence function for the target parameter eq. (8). As a result of this targeting step, the final estimates of the nuisance parameters and of the target parameter satisfy a user-chosen score equation. Under certain regularity conditions, this yields inference based on the Central Limit Theorem. For example, it has been shown that both TMLE and the estimating equation based estimators (using the EIF as estimating equation) are double robust asymptotically efficient estimators under a Donsker class condition and a second order remainder condition [16]. The Donsker class condition roughly states that the EIF at Q_n and g_n falls in a Donsker class such as multivariate real valued cadlag functions with uniformly bounded sectional variation norm. The second order condition states that the product, of (a) the L_2 norm of the estimated outcome regressions minus their true counterpart with (b) the L_2 norm of the estimated treatment/censoring mechanism minus their true counterpart, converges to zero at a rate faster than $n^{1/2}$. The advantage of the TMLE is that it does not rely on the existence of a unique solution to the estimating equation. The TMLE estimators in this section are asymptotically linear and efficient under these conditions. They also allow different combination of rates for Q_n and g_n . For example, if g is known, then only consistency is needed for asymptotic efficiency.

Another advantage of the targeted learning approach is that it does not rely on the existence of a unique solution of the estimating equation. Thus for marginal structural working models that are non-linear in parameters, the non-targeted estimating equations estimators may have multiple solutions or no solution. We refer to [2, 3] for the general methodology.

3.3.1 Efficient influence function

We first establish the EIF for our parameter of interest. From a fundamental result in [17], under standard regularity conditions, the variance of the canonical gradient of Ψ provides a generalized Cramer-Rao lower bound for any regular and asymptotically linear estimators of Ψ . Therefore, this canonical gradient is a vital ingredient in building asymptotically linear and efficient estimators; fittingly, it is also commonly known as the EIF. For parameters in causal inference and missing data applications (such as those in our examples), the EIF also provides insights into the potential robustness against model mis-specifications.

From the first equality in Remark 1, we can obtain the EIF for Ψ via implicit differentiation. We formally state the result here and leave the proof in the Appendix.

Lemma 1

(Efficient Influence Function)

The EIF of Ψ at $P \in \mathcal{M}$ is given by $M(\psi)^{-1}D(\psi)$, where

$$D(\psi) = \sum_{t=0}^K D_t(Q, g) + D_V(Q, g, \psi) + D_W(Q, \psi), \quad (12)$$

with

$$\begin{aligned} D_t(Q, g) &\equiv \sum_{\mathbf{a} \in \mathcal{A}} C_t^{\mathbf{a}}(g) \tilde{h}(\mathbf{a}, V) \{Q_{t+1}^{\mathbf{a}}(\mathbf{L}_t, V, W) - Q_t^{\mathbf{a}}(\mathbf{L}_{t-1}, V, W)\}, \\ D_V(Q, g, \psi) &\equiv \sum_{\mathbf{a} \in \mathcal{A}} \frac{I(A_0 = a_0)}{g(A_0 = a_0 | W)} \{ \tilde{h}(\mathbf{a}, V) [Q_0^{\mathbf{a}}(V, W) - m_{\psi}(\mathbf{a}, V)] \\ &\quad - E_V \{ \tilde{h}(\mathbf{a}, V) [Q_0^{\mathbf{a}}(V, W) - m_{\psi}(\mathbf{a}, V)] | A_0, W \} \}, \\ D_W(Q, \psi) &\equiv \sum_{\mathbf{a} \in \mathcal{A}} E_V \{ \tilde{h}(\mathbf{a}, V) (Q_0^{\mathbf{a}}(V, W) - m_{\psi}(\mathbf{a}, V)) | A_0 = a_0, W \}. \end{aligned}$$

Moreover, if $Q(P) = Q(P_0)$ or $g(P) = g(P_0)$, then $E_{P_0} D(P) = 0$ implies $\Psi(P) = \Psi(P_0)$.

Proof

The proof is given in the Appendix.

3.3.2 Targeted MLE using the EIF $D(\psi)$

The steps described in Section 3.2 can be used to obtain initial estimates of the Q components. The targeted estimators will sequentially update these initial estimates by finding a best fluctuation along a submodel given by components of $D(\psi)$ in eq. (12).

Regarding $Q_t^{\mathbf{a}}$ as a conditional expectation of $Q_{t+1}^{\mathbf{a}}$, as given in eq. (9), we use the quasi loglikelihood loss function for $Q_t^{\mathbf{a}}$:

$$L(Q_t^{\mathbf{a}}) = -\{Q_{t+1}^{\mathbf{a}} \log(Q_t^{\mathbf{a}}) + (1 - Q_{t+1}^{\mathbf{a}}) \log(1 - Q_t^{\mathbf{a}})\}. \quad (13)$$

For a given (Q, g) , and each $t = 0, \dots, K$, consider the d -dimensional working submodel $\{Q_t^{\mathbf{a}}(\epsilon) : \epsilon\}$, with

$$Q_t^{\mathbf{a}}(\epsilon) = \text{expit}(\text{logit} Q_t^{\mathbf{a}} + \epsilon \cdot \tilde{h}(\mathbf{a}, v) C_t^{\mathbf{a}}(g)). \quad (14)$$

This submodel satisfies $\langle \frac{d}{d\epsilon} \sum_{\mathbf{a}} L(Q_t^{\mathbf{a}}(\epsilon)) |_{\epsilon=0} \rangle \supset \langle D_t(Q, g) \rangle$, where $\langle x \rangle$ represents the linear span of a vector x .

Next, we rewrite D_V and D_W in eq. (12) as

$$\begin{aligned} D_V(Q, g, \psi) &\equiv \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \frac{I(A_0 = a_0)}{g(a_0 | W)} \{ \tilde{h}(\mathbf{a}, v) (Q_0^{\mathbf{a}}(V = v, W) - m_{\psi}(\mathbf{a}, v)) \\ &\quad \times (I(V = v) - Q^V(v | A_0, W)) \}, \\ D_W(Q, \psi) &\equiv \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \tilde{h}(\mathbf{a}, v) Q^V(v | A_0 = a_0, W) \{ Q_0^{\mathbf{a}}(V = v, W) - m_{\psi}(\mathbf{a}, v) \}. \end{aligned}$$

We also consider the loss function $L(Q^V) \equiv -\log Q^V$, and a d -dimensional fluctuation model through a given Q^V at $\epsilon = 0$ given by

$$Q^V(\epsilon) \equiv \frac{Q^V(V | A_0 = a_0, W) \exp[\epsilon \cdot B(Q, \psi)]}{\sum_v Q^V(v | A_0 = a_0, W) \exp[\epsilon \cdot B(Q, \psi)]},$$

where $B(Q, \psi) \equiv \sum_{\mathbf{a}} \frac{I(A_0 = a_0)}{g(a_0 | W)} \tilde{h}(\mathbf{a}, V) \{ Q_0^{\mathbf{a}}(V, W) - m_{\psi}(\mathbf{a}, V) \}$. It is easy to verify that $\langle \frac{d}{d\epsilon} L(Q^V(\epsilon)) |_{\epsilon=0} \rangle \supset \langle D_V(Q, g, \psi) \rangle$.

1. Start at $t = K$: regress Y_K on $(\mathbf{L}_{K-1}, \mathbf{A}, L_0, V, W)$, and then evaluate at $\mathbf{A} = \mathbf{a}$ to obtain an initial estimator $\hat{Q}_K^{\mathbf{a}}$ of $Q_K^{\mathbf{a}}$. The optimal fluctuation amount around this initial estimate is given by

$$\hat{\epsilon}_K \equiv \arg \min_{\epsilon} \sum_{\mathbf{a}} P_n L(\hat{Q}_K^{\mathbf{a}}(\epsilon)).$$

This can be implemented by creating one row for each individual and each $\mathbf{a} \in \mathcal{A}$, and fitting a weighted logistic regression of Y_K on the multivariate covariate $\phi(\mathbf{a}, v)C_K^{\mathbf{a}}(\hat{g})$ on these observations with weights $h(\mathbf{a}, v)$ and offset $\hat{Q}_K^{\mathbf{a}}(\mathbf{L}_{K-1}, V, W)$. Update the initial estimator using $\hat{Q}_K^{*,\mathbf{a}} \equiv \hat{Q}_K^{\mathbf{a}}(\hat{\epsilon}_K)$.

- At each subsequent step $t = K - 1, \dots, 0$, we have thus far obtained an updated estimator $\hat{Q}_{t+1}^{*,\mathbf{a}}$ for each $\mathbf{a} \in \mathcal{A}$. Regress $\hat{Q}_{t+1}^{*,\mathbf{a}}$ on $(\mathbf{L}_{t-1}, \mathbf{A}_t, L_0, W, V)$ and evaluate at $\mathbf{A}_t = \mathbf{a}_t$ to get an initial estimator $\hat{Q}_t^{\mathbf{a}}$ of $Q_t^{\mathbf{a}}$. The optimal fluctuation amount around this initial estimator is given by $\hat{\epsilon}_t = \arg \min_{\epsilon} \sum_{\mathbf{a}} P_n L(\hat{Q}_t^{\mathbf{a}}(\epsilon))$, and can be obtained in an analogous manner to step 1. The updated estimator is $\hat{Q}_t^{*,\mathbf{a}} \equiv \hat{Q}_t^{\mathbf{a}}(\hat{\epsilon}_t)$.
- After sequentially performing step (2) in decreasing order of t , we now have a targeted estimator $\hat{Q}_0^{*,\mathbf{a}}$ of $Q_0^{\mathbf{a}}$.
- Let $\hat{Q}^V$ be an estimator of Q^V . For each $\mathbf{a} \in \mathcal{A}$, $v \in \mathcal{V}$ and individual i , create a row of data consisting of $\hat{Q}_0^{*,\mathbf{a}}(V = v, W)$, $h(\mathbf{a}, v)$, $\phi(\mathbf{a}, v)$ and $\hat{Q}^V(v | A_0 = a_0, W)$. Obtain a first-iteration estimator ψ^1 of ψ_0 by fitting a weighted logistic regression of $\hat{Q}_0^{*,\mathbf{a}}(V = v, W)$ onto $\phi(\mathbf{a}, v)$, with weights $h(\mathbf{a}, v) \times \hat{Q}^V(v | A_0 = a_0, W)$, on this pooled data.
- Given ψ^1 obtained in previous step, we update the estimator for Q^V as follows. Using previously obtained $\hat{Q}_0^{*,\mathbf{a}}$, $\hat{g}$, and ψ^1 , the optimal fluctuation amount around the initial $\hat{Q}^V$ is given by $\hat{\epsilon}_V = \arg \min_{\epsilon} P_n \frac{I(A_0=a_0)}{\hat{g}(a_0|W)} L(\hat{Q}^V(\epsilon))$. This can be obtained by solving for ϵ in the equation

$$0 = \sum_{i=1}^n \frac{I((A_0)_i = a_0)}{\hat{g}(a_0 | W_i)} \times \left\{ \hat{B}(W_i, V_i) - \frac{\sum_v \hat{B}(W_i, v) \hat{Q}^V(v | A_0 = a_0, W_i) \exp[\cdot \hat{B}(W_i, v)]}{\sum_v \hat{Q}^V(v | A_0 = a_0, W_i) \exp[\cdot \hat{B}(W_i, v)]} \right\}$$

where $\hat{B} \equiv B((\hat{Q}_0^{*,\mathbf{a}} : \mathbf{a}), \psi^1)$. The updated density is given by $\hat{Q}^{1,V} \equiv \hat{Q}^V(\epsilon^V)$.

- Having obtained a parameter estimate ψ^j and an updated density $\hat{Q}^{j,V}$ at the j -th iteration, repeat step (4) and (5) to obtain a targeted estimate ψ^{j+1} and $\hat{Q}^{j+1,V}$, until ϵ^V converges to 0. In practice, this convergence can be achieved (close to 0) after a few iterations. We denote the final updates as $\psi^{*,TMLE}$ and $\hat{Q}^{*,V}$. We call this estimator, $\psi^{*,TMLE}$, the *full TMLE*.

Let $\hat{Q}^* \equiv (\hat{Q}^W, \hat{Q}^{*,V}, \hat{Q}_0^{*,\mathbf{a}})$, where $\hat{Q}^W$ is the empirical distribution of W . By design, $P_n D^*(\hat{Q}^*, \hat{g}, \psi^{*,TMLE}) = 0$. From Lemma 1, we know that $\psi^{*,TMLE}$ is consistent if either $\hat{Q}^* = Q_0$ or $\hat{g} = g_0$. Under standard regularity and empirical process conditions, $\psi^{*,TMLE}$ is asymptotically linear with influence function $M(\Psi)^{-1} D^*(\Psi)$. The asymptotic covariance of $\sqrt{n}(\psi_n^{*,TMLE} - \psi_0)$ can be estimated by the sample covariance matrix $\Sigma^{*,TMLE}$ of $\{M(\psi^{*,TMLE})^{-1} D^*(\hat{Q}^*, \hat{g}, \psi^{*,TMLE})\}$.

In particular, since $M(\Psi)^{-1} D^*(\Psi)$ is the canonical gradient of Ψ , the estimator $\psi^{*,TMLE}$ is asymptotically efficient if all relevant components in D^* are consistently estimated.

Remark 2

To implement the proposed targeted estimator and satisfy the efficient score equation related to eq. (12), we must specify (and iteratively update) the conditional density of V given A_0, W . When V is high-dimensional, correct specification of these densities may be difficult to accomplish.

As an alternative, a regression-based estimator can be used to directly estimate the conditional expectation $E_V[\tilde{h}(\mathbf{a}, V)(\hat{Q}_0^{*,\mathbf{a}}(V, W) - m_{\psi}(\mathbf{a}, V)) | A_0 = a_0, W]$ in D_V and D_W . A root ψ that solves the corresponding $D_V + D_W$ can be computed via numerical methods. For each ψ in the solution space, one must evaluate the corresponding conditional expectation, for each $\mathbf{a} \in \mathcal{A}$. This can become computationally intractable. Moreover, in such estimating equation based approach, the root ψ to the efficient score equation may not be unique.

If components corresponding to V are not estimated correctly, the resulting parameter estimate would not be efficient, and its consistency would hinge on correct specification of g . In the following section, we trade-off estimation of the V component using a projected influence function that retains most of the robustness properties of the EIF, and the resulting estimation is computationally less demanding as it avoids estimating components of V .

3.3.3 Projected influence function

Lemma 2

(Projected Influence Function)

Consider the setup in Lemma 1. The influence function for Ψ at P under the model $\mathcal{M}_{A_0}$, in which $g(A_0 | W)$ is known, is $M(\psi)^{-1} D^{A_0}(\psi)$, where

$$D^{A_0}(\psi) = \sum_{t=1}^K D_t(\psi) + D_W^{A_0}(Q, g, \psi),$$

(15)

with

$$D_W^{A_0}(Q, g, \psi) \equiv \sum_{\mathbf{a} \in \mathcal{A}} \frac{I(A_0 = a_0)}{g(a_0 | W)} \tilde{h}(\mathbf{a}, V) [Q_1^{\mathbf{a}}(L_0, V, W) - m_{\psi}(\mathbf{a}, V)].$$

In particular, it is a valid estimating function for $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$.

Suppose $g(A_0 | W)$ is correctly specified. If in addition, either $g(A_t | \mathbf{A}_{t-1}, L_{t-1}, V, W)$ or $Q_t^{\mathbf{a}}$ are correct for all $t \geq 1$, then $E_{P_0} D^{A_0}(\Psi) = 0$ implies $\Psi(P) = \Psi(P_0)$.

Proof

The proof is given in the Appendix.

At its face value, the proposed D^{A_0} may seem less robust than the EIF D , as it always relies on consistent estimation of $g(A_0 | W)$. However, as we noted in Remark 2, when V is high-dimensional, there are more standard machine learning algorithms available for estimating $g(A_0 | W)$, which is the propensity score of a binary treatment. Moreover, as it becomes apparent in the next section, the estimators that utilizes D^{A_0} , which we call the *projected TMLE*, are also easier to implement, since standard software packages can be used, and no iterated updates are required in the targeted estimator.

3.3.4 Targeted TMLE using the PIF $D^{A_0}(\psi)$

1. We follow steps (1) and (2) in Section 3.3.2. We use the loss functions eq. (13) and fluctuation models eq. (14) to estimate and update $\hat{Q}_t^{*,\mathbf{a}}$, sequentially from $t = K$ to $t = 1$.
2. We now have a targeted estimator $\hat{Q}_1^{*,\mathbf{a}}$. We obtain an estimator ψ^{pTMLE} , i.e., the *projected TMLE* of ψ_0 by fitting a weighted logistic regression of $\hat{Q}_1^{*,\mathbf{a}}(L_0, V, W)$ on $\phi(\mathbf{a}, V)$, with weights $h(\mathbf{a}, v) \frac{I(A_0 = a_0)}{\hat{g}(a_0 | W)}$.

By construction, $P_n D^{A_0}(\hat{Q}_t^{*,\mathbf{a}}, \hat{g}, \psi^{pTMLE}) = 0$. From Lemma 2, ψ^{pTMLE} is a consistent estimator of ψ_0 if $\hat{g}(A_0 | W)$ is consistent, and either $\hat{Q}_t^{*,\mathbf{a}}$ or $\hat{g}(A_t | \mathbf{A}_{t-1}, L_{t-1}, V, W)$, for $t = 1, \dots, K$ are consistent.

Compared with the *full* TMLE using $D(\psi)$ in the previous section, this estimator is particularly appealing when V is high-dimensional, and still provides more robustness protection than the IPTW or the non-targeted substitution estimators. Moreover, under standard regularity and empirical process conditions, ψ^{pTMLE} is asymptotically linear with influence function $M(\Psi)^{-1} D^{A_0}$. The asymptotic covariance of $\sqrt{n}(\psi^{pTMLE} - \psi_0)$ can be estimated by the sample covariance matrix Σ^{pTMLE} of $M(\psi^{pTMLE})^{-1} D^{A_0}(\hat{Q}_t^{*,\mathbf{a}}, \hat{g}, \psi^{pTMLE})$.

4 Simulation study

In this section, we examine the relative performance of the IPTW estimator (Section 3.1), the G-computation (i.e., non-targeted substitution) estimator (Section 3.2), and the *projected* TMLE estimator (Section 3.3.4) for the parameters of an MSM model. Though the *full* TMLE is the most efficient, as Remark 2 noted, when V is high dimensional, correct specification of all conditional densities of V given A_0, W may be difficult. The simulation is thus focused on the performance of the *projected* TMLE in comparison with the IPTW and G-computation estimators.

4.1 Data generating process and target parameter

We consider an example with data structure $O = (W, A_0, V, L_0, (A_t, L_t) : t = 1, \dots, K)$ with $K = 3$, where A_0 is the missing mechanism for V , A_1 is the treatment assignment, and A_t for $t > 1$ is the indicator of remaining in

the study by time t . Time varying covariate L_t consist of L_t^1, L_t^2 , and the death indicator Y_t . The data generating process is as follows:

$$\begin{aligned}
 (W^1, W^2) &\sim (\text{Bern}(0.3), \text{Bern}(0.7)); \\
 A_0 &\sim \text{Bern}(\text{expit}(1 + 2W^1 + 0.1W^2)); \\
 V \in \{0, 1, 2\} &\sim \{I(V = 1) \sim \text{Bern}(\text{expit}(-2 + 1.2W^1 + 0.7W^2)), \\
 &\quad \{I(V = 2) | V \neq 1\} \sim \text{Bern}(\text{expit}(-0.7 + 1.2W^1 + W^2))\}; \\
 L_0^1 &\sim \text{Bern}(\text{expit}(-0.2 + 2W^1 + 0.5W^2 + 0.2I(V = 1) + 0.4I(V = 2))), \\
 L_0^2 &\sim \text{Bern}(\text{expit}(-0.8 + W^1 + W^2 - 0.3I(V = 1) - 0.1I(V = 2))); \\
 A_t &\sim \begin{cases} \text{Bern}(\text{expit}(-1 + W^1 + 1.3W^2 + 0.1I(V = 1) + 0.1I(V = 2) + 1.2L_0^1 + L_0^2 - 0.7W^1 \times L_0^2 \\ \quad - 0.5W^2 \times L_0^1)), \text{ for } t = 1, \\ \text{Bern}(\text{expit}(2 + W^1 + W^2 + 0.1I(V = 1) + 0.1I(V = 2) + 0.6L_0^1 + 1.2L_0^2 - 0.5A_{t-1} - 0.1t \\ \quad + 0.8L_t^1 - 0.3L_t^2 + 0.1L_{t-1}^1 - 0.2L_{t-1}^2 - 0.3A \times L_0^2 + 0.2A_{t-1} \times W^1 - 0.3A_{t-1} \times L_t^2 \\ \quad - 0.2A_{t-1} \times L_{t-1}^1)), \text{ for } t > 1. \end{cases} \\
 L_t^1 &\sim \text{Bern}(\text{expit}(-1 + W^1 + 0.1W^2 - 0.5I(V = 1) - 0.7I(V = 2) + L_0^1 + 0.3L_0^2 + 1.5A_t + 0.4t \\
 &\quad - L_{t-1}^1 - 0.8A_t \times I(V = 1) - 0.2A_t \times I(V = 2) - 0.3A_t \times W^1)); \\
 L_t^2 &\sim \text{Bern}(\text{expit}(-2 + 0.1W^1 + 0.1W^2 + 0.5I(V = 1) + 0.5I(V = 2) + 0.7L_0^1 + 0.2L_0^2 - A_t + 0.2t \\
 &\quad + L_{t-1}^2 - 0.2A_t \times I(V = 1) - 0.4A_t \times I(V = 2) - 0.3A_t \times L_0^2)); \\
 Y_t &\sim \text{Bern}(\text{expit}(-1.4 + 1.5W^1 + W^2 - 0.7I(V = 1) - 0.8I(V = 2) + L_0^1 + L_0^2 + A_t - L_t^1 - 0.1L_t^2 \\
 &\quad - L_{t-1}^1 - 0.3L_{t-1}^2 + A_t \times I(V = 1) + 0.8A_t \times I(V = 2) - 0.3A_t \times L_t^1 \\
 &\quad - 0.4A_t \times L_{t-1}^1 - 0.3A_t \times L_0^2 - 0.2A_t \times W^1)).
 \end{aligned}$$

Once either the censoring equals to 0 or death process equals to 1, then all subsequent variables are encoded by carrying forward their last observation.

Our interventions of interest are $\mathcal{A} = \{(1, 0, 1, 1), (1, 1, 1, 1)\}$. Under the above distribution, $0.1 < P(A_1 = 1 | \cdot) < 0.95$, and $P(A_t = 1 | \cdot) > 0.5$ for all $t \geq 2$.

We model the dose response $\{\rho(\mathbf{a}, v) : \mathbf{a}, v\}$ by the MSM

$$m_\psi(\mathbf{a}, v) = \text{expit}(\psi_1 + \psi_2 a_1 + \psi_3 v_1 + \psi_4 v_2 + \psi_5 a_1 v_1 + \psi_6 a_1 v_2) = \text{expit}(\psi \cdot \phi(\mathbf{a}, v)),$$

where $\phi(\mathbf{a}, v) = (1, a_1, v_1, v_2, a_1 v_1, a_1 v_2)$, with kernel weights $h(\mathbf{a}, v) = P(a_1 | v, A_0 = a_0)$. Note that in this case, the kernel weights are assumed to be known. The target parameter defined in eq. (8) takes value $\psi_0 = (0.825, 0.105, 0.249, -0.046, 1.474, 0.960)$.

4.2 Estimators

The propensity scores $g(A_0 | W)$ is estimated using Super Learner [15], with base algorithms `glm` and `nnet`, adjusting for W^1 and W^2 . This is considered a correct specification of the propensity for A_0 . We use two estimators for the subsequent propensities at $t \geq 1$: a correctly specified logistic model (shorthand 'gc'), and a misspecified logistic model ('gm') that misses some key baseline and time-varying covariates. The denominator for each $C_t^a(g)$ is truncated below by 0.025. We use two estimators for Q_t^a : the correctly specified estimator ('Qc'), and the misspecified estimator ('Qm') only adjusts for V_1 and V_2 at each time t . Both estimators Super Learner with candidate fitting algorithms `glm` and `nnet`.

We consider three cases of model misspecification on Q_t^a and $g(A_t | \cdot)$ for $t \geq 1$: all correct ('Qc, gc'); correct Q_t^a and misspecified g ('Qc, gm'); misspecified Q_t^a and correct g ('Qm, gc'). For all three cases we always use the same correctly specified $g(A_0 | W)$. We implement the second version of the G-computation estimator in Section 3.2, where the weights are given by g . This estimator changes only under specifications 'Qc, gc' and 'Qm, gc'. The IPTW estimator changes only under specifications 'Qc, gc' and 'Qc, gm'.

4.3 Results

The bias, variance, and coverage probability (for the influence function-based confidence intervals) are appraised using 500 and 2000 repetitions.

In Table 1, we see that when g is misspecified, *projected* TMLE using a correct Q_t^a reduces bias over the misspecified IPTW estimator. Similarly, when Q_t^a is misspecified, *projected* TMLE using the correct g reduces bias over the misspecified G-computation estimator. When comparing the correct vs misspecified G-computation estimators, and the correct vs misspecified IPTW, coefficients involving the adjusted covariates (V_1, V_2) were still estimated very well by the misspecified estimator. Under 'Qc, gc', the correct G-computation estimator converges much slower than the IPTW and the *projected* TMLE estimators. We posit that this may be due to its sole reliance on the nonparametric likelihood estimates. As expected, G-computation has the smallest sample variance, and IPTW has the largest sample variance despite the truncated estimators for g .

Under certain regularity conditions, the IPTW and *projected* TMLE estimators are asymptotically linear. Table 2 tabulates the coverage probability of their influence function based confidence intervals. At the correct models 'Qc,gc', IPTW and *projected* TMLE are asymptotically linear with influence function D^{IPW} and D^{A_0} , respectively. We used $\sqrt{\widehat{\text{var}}D_n^{IPW}/n}$ and $\sqrt{\widehat{\text{var}}D_n^{A_0}/n}$ to estimate their respective standard errors. As sample size grows, the actual coverage probability is quite close to the nominal coverage probability, with IPTW having a better coverage. When one of the components is misspecified, the *projected* TMLE still provides very good coverage, even though theoretically D^{A_0} is only part of its influence curve; we postulate that this is because the influence function based standard error estimates are large relative to the finite sample bias. The misspecified IPTW has very good coverage for the covariates that were adjusted for in the misspecified model, but very bad coverage for the confounded coefficients (A and the intercept).

Table 1: Results: Bias, Variance, MSE for estimators of ψ_0 . Qc = correct Q_t^a , Qm= misspecified Q_t^a , gc=correct g , gm=misspecified g .

ψ	Intercept		A		V_1		V_2		$A \times V_1$		$A \times V_2$	
n	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
Bias												
Qc, gc												
Gcomp	0.334	0.364	0.606	0.603	0.82	0.818	0.560	0.552	1.377	1.390	0.906	0.904
IPW	0.016	0.017	0.056	0.063	0.169	0.131	0.070	0.073	0.402	0.117	0.083	0.099
<i>projected</i> TMLE	0.036	0.001	0.021	0.002	0.078	0.054	0.014	0.008	0.519	0.018	0.002	0.006
Qc, gm												
IPW	0.492	0.512	0.831	0.836	0.094	0.131	0.075	0.072	0.638	0.160	0.026	0.004
<i>projected</i> TMLE	0.022	0.033	0.029	0.028	0.069	0.029	0.02	0.006	0.499	0.015	0.008	0.008
Qm, gc												
Gcomp	0.751	0.773	1.383	1.353	0.66	0.640	0.431	0.414	1.089	1.069	0.752	0.727
<i>projected</i> TMLE	0.01	0.024	0.04	0.064	0.123	0.124	0.048	0.062	0.462	0.103	0.041	0.081
Variance												
Qc, gc												
Gcomp	0.082	0.019	0.112	0.023	0.137	0.033	0.084	0.017	0.038	0.008	0.022	0.005
IPW	0.134	0.035	0.226	0.051	0.939	0.118	0.353	0.068	9.605	0.235	0.555	0.112
<i>projected</i> TMLE	0.113	0.029	0.187	0.041	0.734	0.091	0.279	0.053	9.149	0.169	0.398	0.092
Qc, gm												
IPW	0.114	0.030	0.196	0.045	0.897	0.093	0.259	0.056	9.298	0.184	0.418	0.094
<i>projected</i> TMLE	0.097	0.026	0.16	0.036	0.718	0.078	0.21	0.048	9.062	0.141	0.319	0.078
Qm, gc												
Gcomp	0.095	0.024	0.146	0.035	0.147	0.045	0.091	0.024	0.191	0.057	0.071	0.025
<i>projected</i> TMLE	0.123	0.031	0.201	0.044	0.839	0.105	0.316	0.057	9.424	0.207	0.464	0.096

MSE

Qc, gc												
Gcomp	0.195	0.151	0.479	0.387	0.809	0.701	0.397	0.322	1.934	1.941	0.842	0.822
IPW	0.134	0.035	0.229	0.054	0.967	0.135	0.358	0.073	9.767	0.248	0.562	0.122
projected TMLE	0.115	0.029	0.187	0.041	0.740	0.094	0.279	0.053	9.419	0.17	0.398	0.092
Qc, gm												
IPW	0.357	0.292	0.887	0.744	0.906	0.111	0.264	0.062	9.704	0.21	0.419	0.094
projected TMLE	0.097	0.027	0.16	0.036	0.723	0.079	0.210	0.048	9.311	0.141	0.319	0.078
Qm, gc												
Gcomp	0.66	0.621	2.059	1.865	0.583	0.455	0.277	0.195	1.377	1.2	0.6378	0.554
projected TMLE	0.123	0.031	0.203	0.048	0.854	0.12	0.318	0.061	9.637	0.218	0.466	0.103

Gcomp: G-computation estimator. IPW: inverse probability weighted estimator. TMLE: targeted maximum likelihood estimator.

Table 2: Coverage Probability for the Asymptotically Linear Estimators, using Influence-Function based Confidence Interval. Qc = correct Q_t^a , Qm= misspecified Q_t^a , gc=correct g , gm=misspecified g .

n	Intercept		A		V_1		V_2		$A \times V_1$		$A \times V_2$	
	500	2000	500	2000	500	2000	500	2000	500	2000	500	2000
Qc, gc												
IPW	0.948	0.946	0.944	0.940	0.966	0.930	0.940	0.952	0.912	0.950	0.956	0.960
projected TMLE	0.934	0.932	0.934	0.942	0.936	0.924	0.918	0.946	0.906	0.942	0.942	0.938
Qc, gm												
IPW	0.698	0.146	0.556	0.022	0.956	0.928	0.958	0.958	0.928	0.942	0.958	0.964
projected TMLE	0.964	0.956	0.954	0.956	0.984	0.984	0.972	0.986	0.932	0.954	0.966	0.964
Qm, gc												
projected TMLE	0.942	0.930	0.942	0.948	0.956	0.930	0.936	0.954	0.896	0.942	0.954	0.958

IPW: inverse probability weighted estimator.

5 Data analysis example

To illustrate the application of the *projected* TMLE, we revisit our STAR*D example. After an initial screening, patients are enrolled into level 1 of the treatment, where everyone is treated with citalopram. At the start of each subsequent level, if a patient still remains in the study, then he/she is randomized to an option (i.e. a particular medication) within one of his/her accepted treatment strategies (augmenting vs switching). Regular follow-up visits are conducted throughout each level. At each follow-up visit, covariates are collected, and the patient is subject to dropout, entering remission, or moving onto next level if treatments fail.

In the field of depression research, there is very little consistent literature concerning individual baseline characteristics that may differentially modify the effect of switching medication vs augmenting medication on remission. Because the strategies themselves are not randomized, this analysis is *de facto* equivalent to that of an observational study. Our estimators will account for baseline and time-varying confounding. All candidate modifiers are *a priori* selected through a literature review of previous STAR*D publications. These covariates are measured prior to the assignment of level 2 treatment, either at study screening, clinical visits or exit surveys at the end of level 1. Many of these candidate modifiers are subject to missingness, often times due to missing visits and surveys, therein lies the need for the tools developed here. Our study population is the set of 1395 patients at level 2 who found medication strategies acceptable. Table 3 tabulates the events in level 2 by strategies received. Note that there are three strategies, but we are only comparing switching medication vs augmenting medication. The data structure was described in detail in Section 2 as part of our running example.

We consider here two types of potential effect modifiers: some are measured at enrollment and some are measured at exit of level 1. If V is a baseline covariate, W includes all demographic variables and medical and

psychiatric history prior to enrollment; if V is a level 1 exit covariate, then we add to W variables summarizing number of visits, adherence to study protocol, and time spent in level 1. Table 4 summarizes percent of missingness, range, and scale of each effect modifier. Of all the candidate modifiers proposed by literature review, we exclude from our analysis the history of amphetamine use at baseline and history of drug abuse, since these two variables are missing for more than 70% of the patients. The multivariate nature of most of the effect modifiers underscores the need for *projected* TMLE.

Table 3: Level 2 events by acceptable strategy received.

Received	Medication-Sw	Medication-Aug	Cognitive Therapy
n	727	565	103
Remission	257 (35%)	287 (51%)	53 (51%)
No Remission	227 (32%)	132 (23%)	22 (21%)
Dropout	243 (33%)	146 (26%)	28 (27%)

The analysis below compares medication switching (Sw) and medication augmentation (Aug) only.

Table 4: Candidate effect modifiers of interest for level 2. Candidates with more than 50% missing in the study population will be discarded from the analysis.

	% missing	Type (# of categories, or range of semi-continuous variable)
At screening		
Race	0	Categorical (3)
Years of schooling completed	0	Semi-continuous (0,27)
Employment status	0	Categorical (3)
Education level	0	Categorical (3)
Attempted suicide	0	Categorical (2)
Duration of current major depression episode (in units of 6 months)	1	Semi-continuous (0,111)
AxI: Amphetamine	73 (will not use)	Categorical (2)
History of drug abuse	76 (will not use)	Categorical (2)
# of Cumulative Illness Rating (CIRS) categories	0	Semi-continuous (0,12)
CIRS severity score	0	Semi-continuous (0,4)
Have AxI Obsessive-Compulsive Disorder	1	Categorical (2)
Have AxI Panic Disorder	1	Categorical (2)
Have AxI Post Traumatic Stress Disorder	2	Categorical (2)
Have AxI Agoraphobia	1	Categorical (2)
Have AxI Social Anxiety Disorder	2	Categorical (2)
Have AxI Generalized Anxiety Disorder	2	Categorical (2)
Anxious Depression at screening	6	Categorical (2)
Atypical Depression at screening	5	Categorical (2)
Melancholic Depression at screening	5	Categorical (2)
At level 1 exit		
Maximum Side Effects	1	Categorical (7)
SF-36 Physical Health (in units of 5)	23	Semi-continuous (0,12)
SF-36 Mental Health (in units of 5)	23	Semi-continuous ((0,12)
Quality of Life	23	Semi-continuous (0,98)
Side effect burden	1	Categorical (7)
Quick Inventory of Depressive Symptom (QIDS) total score (self-reported)	1	Semi-continuous (0,27)
Hamilton rating scale for depression (17-time) score	12	Semi-continuous (0,43)
IDS-C (30-item) score	13	Semi-continuous (0,74)

AxI: axis I disorders. CIRS: Cumulative Illness Rating Scale. IDS-C: 30-item Inventory of Depressive Symptomatology — Clinician Rated.

The MSM is a generalized linear model with logit link. The linear predictor $\phi(\mathbf{a}, v) = (1, a_1, v_1, \dots, v_k, a_1 \times v_1, \dots, a_1 \times v_k)$ and $\psi \equiv (\psi_1, \psi_2, \psi_{3,1}, \dots, \psi_{3,k}, \psi_{4,1}, \dots, \psi_{4,k}) \in \mathbb{R}^{2k+2}$ for the categorical V with $k+1$ categories, and $\phi(\mathbf{a}, v) = (1, a_1, v, v^2, a_1 \times v, a_1 \times v^2)$ and $\psi \equiv (\psi_1, \psi_2, \psi_{3,1}, \psi_{3,2}, \psi_{4,1}, \psi_{4,2}) \in \mathbb{R}^6$ for the semi-continuous V . The kernel weights are $h(\mathbf{a}, v) = P(A_1 = a_1 | V)$, to be estimated using Super Learner with fitting algorithms `glm`, `knnreg`, `nnet` and `bayesglm`. The Super Learner will use 10-fold cross validation to select the best weighted combination of these fitting algorithms to estimate $h(\mathbf{a}, v)$. The initial estimators of g and Q_t^a adjust for all baseline covariates and time-varying covariates with up to 2 time lag (each covariate is coupled with its missingness indicator). Each fitting algorithm is coupled with each of the following screening algorithms: Spearman correlation tests at significance levels 0.01, 0.05, 0.1, 0.2; ranking p-values from the correlation tests and take the top k variables, where k ranges from 10 variables up to 30% of the total number of variables being considered, in increments of 20. The Super Learner selects the best linear combination of all screening-fitting couples.

The goal of the analysis is to identify potential effect modifiers from a pool of candidate pre-treatment covariates. Our strategy is to measure the treatment heterogeneity for each of these covariates using an MSM, and then apply multiple testing methods to identify those for which the treatment heterogeneity is significant. A common way to assess treatment heterogeneity across strata of V is to test whether the coefficients of the interaction terms ψ_4 are different from 0. We perform the corresponding Wald test in Table 5, with the null hypothesis $\psi_4 = 0$ and test statistics $T_n = \psi_{4,n}^\top \Sigma_n^{-1} \psi_{4,n} \sim \chi_k^2$, where $\psi_{4,n}$ is the *projected* TMLE estimator, $\Sigma_n = \widehat{Cov}(D_{\psi_4}^{A_0})/n$, and k is the number of interaction terms. The false discovery rate (FDR) of the simultaneous comparisons is controlled at 0.05 with the Benjamini-Hochberg procedure.

Table 5: Data Analysis Results: Multivariate Wald Test. $\psi_4 = (\psi_{4,1}, \dots, \psi_{4,k})$ are the coefficients of the interaction terms in the MSM; if V is semi-continuous, interaction terms are $\psi_4 \cdot (v, v^2)$, if V is categorical with $k+1$ categories, interaction terms are $\psi_{4 \cdot (v_1, \dots, v_k)}$. $H_0: \psi_4 = 0$. $T_n = \psi_{4,n}^\top \Sigma_n^{-1} \psi_{4,n} \sim \chi_k^2$, where $\psi_{4,n}$ is given by *projected* TMLE, and Σ_n is the corresponding covariance estimator based on the covariance matrix of the projected-IF. Control false discovery rate at 0.05 with the Benjamini-Hochberg procedure.

	P-value	Reject	ψ_4	$SE(\psi_4)$
At screening				
Race	4.97e-01	F	(-1.68e-01, -4.62e-01)	(2.62e-01, 4.27e-01)
Years of schooling completed	2.89e-01	F	(2.50e-01, -1.10e-02)	(1.71e-01, 7.07e-03)
Employment status	1.50e-01	F	(3.92e-01, -3.94e-01)	(2.30e-01, 5.89e-01)
Education level	1.13e-01	F	(4.89e-01, -5.69e-03)	(3.13e-01, 3.80e-01)
Attempted suicide	1.80e-05	T	(-1.10e+00)	(2.56e-01)
Duration of current major depression episode (in units of 6 months)	4.44e-02	F	(1.92e-02, -9.41e-04)	(2.78e-02, 5.59e-04)
# of Cumulative Illness Rating (CIRS) categories	2.05e-02	F	(3.40e-01, -4.51e-02)	(1.42e-01, 1.65e-02)
CIRS severity score	8.09e-02	F	(-6.21e-01, 3.65e-01)	(4.41e-01, 1.85e-01)
Have AxI Obsessive-Compulsive Disorder	5.20e-01	F	(-1.86e-01)	(2.89e-01)
Have AxI Panic Disorder	1.17e-01	F	(5.53e-01)	(3.53e-01)
Have AxI Post-Traumatic Stress Disorder	8.78e-01	F	(3.86e-02)	(2.51e-01)
Have AxI Agoraphobia Disorder	9.73e-01	F	(1.04e-02)	(3.12e-01)
Have AxI Social Anxiety Disorder	5.15e-01	F	(1.81e-01)	(2.78e-01)
Have AxI Generalized Anxiety Disorder	1.90e-01	F	(3.68e-01)	(2.81e-01)
Anxious Depression at screening	6.89e-01	F	(-9.17e-02)	(2.3e-01)
Atypical Depression at screening	4.15e-04	T	(9.79e-01)	(2.77e-01)
Melancholic Depression at screening	1.58e-01	F	(-3.48e-01)	(2.46e-01)
At level 1 exit				
Maximum side effects	6.46e-09	T	(5.53e-01, 8.10e-02, 1.85e+00, -2.62e-01, 1.39e+00, 5.88e-01)	(5.27e-01, 5.25e-01, 4.59e-01, 5.02e-01, 5.20e-01, 5.65e-01)
SF-36 Physical Health (in units of 5)	6.83e-09	T	(2.55e+00, -1.74e-01)	(4.17e-01, 2.91e-02)
SF-36 Mental Health (in units of 5)	1.36e-01	F	(5.55e-01, -5.48e-02)	(2.78e-01, 2.86e-02)
Quality of Life	4.84e-04	T	(-7.53e-02, 5.35e-04)	(3.12e-02, 3.46e-04)

Side effect burden	0.00e+00	T	(2.91e−01, 5.18e−01, −6.26e−02, 1.29e+00, 2.73e+00, −5.85e−01)	(2.00e−03, 2.41e−03, 2.60e−03, 2.50e−03, 1.96e−03, 1.21e−03)
Quick Inventory of Depressive Symptom (QIDS) to- tal score (self-reported)	7.61e−02	F	(2.82e−01, −1.03e−02)	(1.24e−01, 4.62e−03)
Hamilton rating scale for depression (17-item) score	4.83e−02	F	(1.16e−01, −1.88e−03)	(8.45e−02, 2.17e−03)
IDS-C (30 item) score	1.27e−01	F	(5.06e−02, −4.75e−04)	(4.47e−02, 6.72e−04)

AxI: axis I disorders. IDS-C: 30-item Inventory of Depressive Symptomatology — Clinician Rated.

Since these candidate covariates differ in their type (semi-continuous vs binary vs multiple categories), the corresponding MSM parameters are of different dimensions, we also consider the measure of treatment heterogeneity

$$\beta(\psi) \equiv \max_v \ell_{OR}(v; \psi) - \min_v \ell_{OR}(v; \psi),$$

where

$$\ell_{OR}(v; \psi) = \log \frac{m(1, v; \psi)}{1 - m(1, v; \psi)} - \log \frac{m(0, v; \psi)}{1 - m(0, v; \psi)} = \begin{cases} \psi_2 + \psi_{4,1}v_1 + \cdots + \psi_{4,k}v_k, & \text{for categorical } V, \\ \psi_2 + \psi_{4,1}v + \psi_{4,2}v^2, & \text{for semi-continuous } V. \end{cases}$$

This measure quantifies the most change in log odds ratio between any two values of V . Consider the null hypothesis $H_0 : \beta(\psi_0) = 0$. Using *projected* TMLE, we obtain an estimator $\beta_n = \beta(\psi_n^{pTMLE})$ for each V . An application of the functional delta method, with the covariance matrix Σ_n of the estimated influence function $D_n^{A_0}$, yields a standard error estimate SE_n for β_n . We use the test statistics $T_n = \beta_n / SE_n \sim N(0, 1)$. The results of the analysis are summarized in Table 6. Interestingly, this approach yields almost all the same potential effect modifiers as in Table 5.

Using our methods, we identified six factors, two at screening and four at level 1 exit, that modified the comparative treatment effect of augmenting vs switching medication. These effect modifiers are: prior suicide attempts, atypical depression, level 1 side-effect frequencies and burden (impairment), SF-36 physical functionality, and quality of life at level 1 exit. In general, the heterogeneity of the treatment effects are most pronounced for the level 1 exit modifiers than for those from baseline, indicating opportunities for close monitoring and medication management.

There has only been one publication that addressed the heterogeneous treatment effects between switching and augmenting in level 2 treatment in STAR*D [18]. The authors employed propensity score and weighting methods to examine heterogeneity of the treatment effect among the treated (medication augmentation) through stratification by propensity score decile. This approach does not explicitly identify factors contributing to heterogeneity. Missing data is less problematic in this case, and is handled by multiple imputation.

6 Summary

In this work, we studied causal effect modification by a counterfactual modifier. The tools developed here are applicable in situations where the effect modifier of interest may be better cast as counterfactual variables. Examples of such situations include the study of time-varying effect modification, or the study of baseline effect modification with modifiers that are missing at random.

We established the efficient influence function (EIF) for the corresponding marginal structural model parameters which provides the semiparametric efficiency bound for all asymptotically linear estimators. This EIF is also doubly robust in that it remains an unbiased estimating function if either 1) the outcome expectations and the modifier densities, or 2) the treatment and censoring mechanisms, are consistently estimated. However, in applications with high-dimensional effect modifiers, we saw that it may be difficult to fully utilize the EIF without potentially compromising consistency. To solve this problem, we presented a projected influence function (PIF), which equals the EIF in a model where the missingness mechanism (or more generally, pre-modifier intervention, A_0) is known. Though not fully efficient under the larger model, this PIF is robust against misspecification of the outcome models or the treatment mechanisms, whenever the missingness mechanism is consistently estimated. We presented two TMLE estimators using the EIF and the PIF. We also described an IPTW estimator that is unbiased if the intervention probabilities are consistently estimated, and a non-targeted G-computation estimator that is unbiased if the outcome expectations and either the modifier densities or the

missingness mechanisms are consistently estimated. Under standard regularity and empirical process conditions, the two TMLE estimators and the IPTW estimators are asymptotically linear, thereby allowing Central Limit Theorem based standard error estimates. Moreover, the full TMLE estimator using the EIF will be semi-parametric efficient if all the components of the likelihood are consistently estimated.

Table 6: Data Analysis Results: $\beta_0 = \max_v \ell_{OR}(v; \psi_0) - \min_v \ell_{OR}(v; \psi_0)$, $H_0 : \beta_0 = 0$. $T_n = \beta_n / SE_n \sim N(0, 1)$, where ψ_n given by *projected* TMLE. The standard error estimate $\hat{SE}(\beta_n)$ is computed using the variance of the projected influence function and the functional delta method. Control false discovery rate at 0.05 with the Benjamini-Hochberg procedure.

	P	Reject	β_n	$\hat{SE}(\beta_n)$	$\arg \min_v \ell_{OR}(v; \psi_0)$	$\arg \max_v \ell_{OR}(v; \psi_0)$
At screening						
Race	2.8e−01	F	4.6e−01	4.3e−01	3.0e+00	2.0e+00
Years of schooling completed	1.2e−01	F	2.7e+00	1.7e+00	2.7e+01	1.1e+01
Employment status	1.9e−01	F	7.9e−01	6.0e−01	3.0e+00	2.0e+00
Education level	9.2e−02	F	4.9e−01	2.9e−01	3.0e+00	2.0e+00
Attempted suicide	1.8e−05	T	1.1e+00	2.6e−01	1.0e+00	0.0e+00
Duration of current major depression episode (in units of 6 months)	3.3e−02	F	9.6e+00	4.5e+00	1.1e+02	1.0e+01
# of Cumulative Illness Rating (CIRS) categories	5.9e−03	T	3.1e+00	1.1e+00	1.2e+01	3.8e+00
CIRS severity score	2.7e−02	F	3.6e+00	1.6e+00	8.5e−01	4.0e+00
Have AxI Obsessive-Compulsive Disorder	5.2e−01	F	1.9e−01	2.9e−01	1.0e+00	0.0e+00
Have AxI Panic Disorder	1.2e−01	F	5.5e−01	3.5e−01	0.0e+00	1.0e+00
Have AxI Post-Traumatic Stress Disorder	8.8e−01	F	3.9e−02	2.5e−01	0.0e+00	1.0e+00
Have AxI Agoraphobia Disorder	9.7e−01	F	1.0e−02	3.1e−01	0.0e+00	1.0e+00
Have AxI Social Anxiety Disorder	5.2e−01	F	1.8e−01	2.8e−01	0.0e+00	1.0e+00
Have AxI Generalized Anxiety Disorder	1.9e−01	F	3.7e−01	2.8e−01	0.0e+00	1.0e+00
Anxious Depression at screening	6.9e−01	F	9.2e−02	2.3e−01	1.0e+00	0.0e+00
Atypical Depression at screening	4.2e−04	T	9.8e−01	2.8e−01	0.0e+00	1.0e+00
Melancholic Depression at screening	1.6e−01	F	3.5e−01	2.5e−01	1.0e+00	0.0e+00
At level 1 exit						
Maximum side effects	8.1e−09	T	2.1e+00	3.7e−01	4.0e+00	3.0e+00
SF-36 Physical Health (in units of 5)	9.0e−10	T	9.3e+00	1.5e+00	0.0e+00	7.3e+00
SF-36 Mental Health (in units of 5)	1.0e−01	F	2.6e+00	1.6e+00	1.2e+01	5.1e+00
Quality of Life	1.4e−04	T	2.6e+00	7.0e−01	7.0e+01	0.0e+00
Side effect burden	0.0e+00	T	3.3e+00	2.3e−03	6.0e+00	5.0e+00
Quick Inventory of Depressive Symptom (QIDS) to- tal score (self-reported)	2.7e−02	F	1.9e+00	8.7e−01	0.0e+00	1.4e+01
Hamilton rating scale for depression (17-item) score	2.0e−02	F	1.8e+00	7.6e−01	0.0e+00	3.1e+01
IDS-C (30 item) score	5.1e−02	F	1.3e+00	6.9e−01	0.0e+00	5.3e+01

AxI: axis I disorders. IDS-C: 30-item Inventory of Depressive Symptomatology — Clinician Rated.

Acknowledgements

We thank Josh Schwab for his invaluable help in implementing the *projected* TMLE.

Funding

This work was supported by NIH grants RC4MH092737 (Principal Investigator Z. Luo) and R01 A1074345 (Principal Investigator M. van der Laan).

Appendix

A Proof of identifiability

$$\begin{aligned}
 P(V(a_0) = v) &= \sum_w P(w, V(a_0) = v) = \sum_w P(w)P(V(a_0) = v | w) \\
 &\stackrel{\text{SRA}}{=} \sum_w P(w)P(V = v | A_0 = a_0, W = w) = E_W \{P(V = v | A_0 = a_0, W)\} \equiv \gamma(a_0, v). \\
 E[Y_K(\mathbf{a}) | V(a_0) = v] &= \sum_{w, \mathbf{l}} y_K \frac{P(\mathbf{L}(\mathbf{a}) = \mathbf{l}, L_0(a_0) = l_0, V(a_0) = v, W = w)}{P(V(a_0) = v)} \\
 &\stackrel{\text{SRA}}{=} \sum_{w, \mathbf{l}} y_K \frac{P(w)P(V = v | A_0 = a_0, w)P(l_0 | w, A_0 = a_0, V = v) \left(\prod_{t=1}^K P(l_t | \mathbf{A}_t = \mathbf{a}_t, \mathbf{l}_{t-1}, V = v, A_0 = a_0, w) \right)}{P(V(A_0 = a_0) = v)} \\
 &= \frac{1}{\gamma(P)(a_0, v)} E_P \{P(V = v | A_0 = a_0, W) Q_0^{\mathbf{a}}(V = v, W)\}
 \end{aligned}$$

B Proof of Remark 1

The first equality in the remark follows from definition of $\Psi(P)$ and choice of $m_\psi(\mathbf{a}, v)$, the third equality is trivial. We only show the second one.

$$\begin{aligned}
 &\sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \tilde{h}(\mathbf{a}, v) \{ \eta(P)(\mathbf{a}, v) - \gamma(P)(a_0, v) m_{\Psi(P)}(\mathbf{a}, v) \} \\
 &= \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \tilde{h}(\mathbf{a}, v) (E_W p(V = v | a_0, W) Q_0^{\mathbf{a}}(v, W) - E_W (p(V = v | a_0, W)) m_{\Psi(P)}(\mathbf{a}, v)) \\
 &= E_W \left\{ \sum_{\mathbf{a}, v} p(V = v | a_0, W) \tilde{h}(\mathbf{a}, v) (Q_0^{\mathbf{a}}(v, W) - m_{\Psi(P)}(\mathbf{a}, v)) \right\}
 \end{aligned}$$

The first-line in the right-hand-side of the above equation is zero by definition of $\eta(P)(\mathbf{a}, v)$.

C Proof of Lemma 1: Efficient influence function for $\Psi(P)$ under $\mathcal{M}$

C In this appendix, we derive the EIF at P of the map $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$. For each $P \in \mathcal{M}$, let $\mathcal{H}_P$ denote the Hilbert space of 1-dimensional mean zero measurable functions of O with finite variance, endowed with the covariance inner product. For an $r \in \mathcal{H}_P$, define a 1-dimensional parametric submodel $\{P_r(\alpha) : |\alpha| < 1/\|r\|_\infty\}$, through P at $\alpha = 0$, given by $\frac{dP_r(\alpha)}{d\alpha} = 1 + \alpha r(O)$. Since we are working under a saturated model $\mathcal{M}$, this submodel is indeed contained in $\mathcal{M}$. We shall consider the class of all such 1-dimensional submodels indexed by $\mathcal{H}_P$.

For a given $\tilde{D} \in \mathcal{H}_P^d$, we define the vector inner product $E_P(\tilde{D} \times r)$ as the vector of the component-wise inner products $(E_P(\tilde{D}_j \times r) : j = 1 \dots d)$. We wish to show that D satisfies

$$\left. \frac{d\Psi(P_r(\alpha))}{d\alpha} \right|_{\alpha=0} = M(\Psi(P), P)^{-1} E_P(D \times r).$$

From definition of the maps Ψ , $\gamma(a_0, v)$ and $\eta(\mathbf{a}, v)$, and our choice of working model $m_\psi(\mathbf{a}, v)$, we know that at each $P_r(\alpha)$

$$0 = \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \gamma(P_r(\alpha))(a_0, v) h(\mathbf{a}, v) \phi(\mathbf{a}, v) \left[\frac{\eta(P_r(\alpha))(\mathbf{a}, v)}{\gamma(P_r(\alpha))(a_0, v)} - m_{\Psi(P_r(\alpha))}(\mathbf{a}, v) \right]. \quad (16)$$

We perform an implicit differentiation with respect to α on the above equation, at $\alpha = 0$, to obtain the equality

$$\begin{aligned} & \left. \frac{d\Psi(P_r(\alpha))}{d\alpha} \right|_{\alpha=0} \times \\ & \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \gamma(P)(a_0, v) h(\mathbf{a}, v) \phi(\mathbf{a}, v) m_{\Psi(P)}(\mathbf{a}, v) [1 - m_{\Psi(P)}(\mathbf{a}, v)] \phi(\mathbf{a}, v)^\top \right\} \\ & = \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) \left\{ \left. \frac{d\eta(P_r(\alpha))(\mathbf{a}, v)}{d\alpha} \right|_{\alpha=0} - \left(\left. \frac{d\gamma(P_r(\alpha))(a_0, v)}{d\alpha} \right|_{\alpha=0} \right) m_{\Psi(P)}(\mathbf{a}, v) \right\}. \end{aligned} \quad (17)$$

Next, for a given $(\mathbf{a}, v)$, we proceed to express $\left. \frac{d\gamma(P_r(\alpha))}{d\alpha} \right|_{\alpha=0}$ and $\left. \frac{d\eta(P_r(\alpha))}{d\alpha} \right|_{\alpha=0}$ as $E_P(D_\gamma \times r)$ and $E_P(D_\eta \times r)$, respectively, for some functions D_γ and D_η belonging to the Hilbert space $\mathcal{H}_P$.

For convenience of indexing, for a given vector $\mathbf{l}_K$, we shall use the short hand $h_t^{\mathbf{a}, v} = (A_0 = a_0, V = v, \mathbf{L}_{t-1} = \mathbf{l}_{t-1}, \mathbf{A}_t = \mathbf{a}_t)$, for $t = 1, \dots, K$, and $h_t^{\mathbf{a}, v} = (A_0 = a_0, V = v)$ for $t = 0$. From definition of $P_r(\alpha)$, it follows that

$$\begin{aligned} P_r(\alpha)(W) &= P(W) (1 + \alpha E_P(r | W)), \\ P_r(\alpha)(v | A_0 = a_0, W) &= P(v | A_0 = a_0, W) \frac{1 + \alpha E_P(r | V = v, A_0 = a_0, W)}{1 + \alpha E_P(r | A_0 = a_0, W)}, \\ \text{and } P_r(\alpha)(l_t | W, H_t^{\mathbf{a}, v}) &= P(l_t | W, H_t^{\mathbf{a}, v}) \frac{1 + \alpha E_P(r | l_t, H_t^{\mathbf{a}, v})}{1 + \alpha E_P(r | H_t^{\mathbf{a}, v})}. \end{aligned}$$

Proposition 1

For a given (a_0, v) , $\left. \frac{d\gamma(P_r(\alpha))}{d\alpha} \right|_{\alpha=0} = E_P(D_\gamma \times r)$, where

$$D_\gamma(P) = \frac{I(A_0 = a_0)}{g(a_0 | W)} (I(V = v) - P(v | A_0 = a_0, W)) + P(v | A_0 = a_0, W) - \gamma(P)(a_0, v). \quad (18)$$

Proof

$$\begin{aligned} \left. \frac{d\gamma(P_r(\alpha))}{d\alpha} \right|_{\alpha=0} &= \lim_{\alpha \rightarrow 0} \frac{\gamma(P_r(\alpha)) - \gamma(P)}{\alpha} = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} \int_W \left\{ P(W) P(v | A_0 = a_0, W) \alpha \right. \\ & \quad \left. \frac{E_P(r | v, A_0 = a_0, W) - E_P(r | A_0 = a_0, W) + E_P(r | W) + \alpha E_P(r | W) E_P(r | v, A_0 = a_0, W)}{1 + \alpha E_P(r | A_0 = a_0, W)} \right\} \\ &= \int_W P(W) P(v | A_0 = a_0, W) \{ E_P(r I(V = v) | A_0 = a_0, W) - E_P(r | A_0 = a_0, W) + E_P(r | W) \} \\ &= E_P \left\{ \left(\frac{I(A_0 = a_0)}{g(a_0 | W)} (I(V = v) - P(v | A_0 = a_0, W)) + P(v | A_0 = a_0, W) \right) \times r(O) \right\} \\ &= E_P(D_\gamma(P) \times r(O)). \end{aligned}$$

In obtaining the last equality, we note that centering the left factor of the integrand by $\gamma(P)(a_0, v)$ does not change the expression because $E_P(\gamma(P)r(O)) = \gamma(P)E_P(r(O)) = 0$ by definition of r . It is straightforward to check that indeed $E_P D_\gamma(P) = 0$. Moreover, under our saturated model $D_\gamma(P)$ is in fact in the tangent space. This concludes the proof of Proposition 1.

Proposition 2

For a given $\mathbf{a} \in \mathcal{A}$, $v \in \mathcal{V}$, $\left. \frac{d\eta(P_r(\alpha))}{d\alpha} \right|_{\alpha=0} = E_P(D_\eta \times r)$, where

$$\begin{aligned} D_\eta(P) &\equiv I(V = v) \sum_{t=0}^K C_t^{\mathbf{a}} \{ Q_{t+1}^{\mathbf{a}}(\mathbf{L}_t, v, W) - Q_t^{\mathbf{a}}(\mathbf{L}_{t-1}, v, W) \} \\ &+ \frac{I(A_0 = a_0)}{g(a_0 | W)} Q_0^{\mathbf{a}}(v, W) \{ I(V = v) - P(v | A_0 = a_0, W) \} \\ &+ P(v | A_0 = a_0, W) Q_0^{\mathbf{a}}(v, W) - \eta(P)(\mathbf{a}, v). \end{aligned}$$

$$\text{where } C_t^a = \frac{I(\mathbf{A}_t = \mathbf{a}_t)}{\prod_{j=0}^t g(a_j | A_{j-1} = a_{j-1}, L_{j-1}, V, W)}.$$

Proof

$$\begin{aligned} \left. \frac{d\eta(P_r(\alpha))(\mathbf{a}, v)}{d\alpha} \right|_{\alpha=0} &= \lim_{\alpha \rightarrow 0} \frac{\eta(P_r(\alpha))(\mathbf{a}, v) - \eta(\mathbf{a}, v)(P)}{\alpha} \\ &= \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} \int_{w, l_0, \mathbf{l}_K} y_K P(w) P(v | A_0 = a_0, w) \prod_{j=0}^K P(l_j | w, h_j^{\mathbf{a}, v}) \alpha \\ &\quad \times \left\{ \frac{\sum_{t=0}^K E_P(r | l_t, w, h_t^{\mathbf{a}, v}) + E_P(r | v, A_0 = a_0, W) + E_P(r | W)}{1 + \alpha M'_P(w, l_0, \mathbf{l}_K)} \right. \\ &\quad \left. - \frac{\sum_{t=0}^K E_P(r | w, h_t^{\mathbf{a}, v}) - E_P(r | A_0 = a_0, W) + \alpha M_P(w, l_0, \mathbf{l}_K)}{1 + \alpha M'_P(w, l_0, \mathbf{l}_K)} \right\} \\ &= \int_{w, l_0, \mathbf{l}_K} y_K P(w) P(v | A_0 = a_0, w) \prod_{j=0}^K P(l_j | w, h_j^{\mathbf{a}, v}) \\ &\quad \times \left\{ \sum_{t=0}^K E_P(r | l_t, w, h_t^{\mathbf{a}, v}) - \sum_{t=0}^K E_P(r | w, h_t^{\mathbf{a}, v}) + E_P(r | v, A_0 = a_0, w) \right. \\ &\quad \left. - E_P(r | A_0 = a_0, w) + E_P(r | w) \right\} \\ &= E_P \left\{ \frac{I(A_0 = a_0)}{g(a_0 | W)} I(V = v) \sum_{t=1}^K \frac{I(\mathbf{A}_t = \mathbf{a}_t)}{\prod_{j=1}^t g(a_t | \text{parents}(a_t))} \right. \\ &\quad \times [Q_{t+1}^{\mathbf{a}}(\mathbf{l}_t, v, w) - Q_t^{\mathbf{a}}(\mathbf{l}_{t-1}, v, w)] \times r(O) \Big\} \\ &\quad + E_P \left\{ \frac{I(A_0 = a_0)}{g(a_0 | W)} Q_0^{\mathbf{a}}(V = v, W) [I(V = v) - P(v | A_0 = a_0, W)] \times r(O) \right\} \\ &\quad + E_P \{ P(v | A_0 = a_0, W) Q_0^{\mathbf{a}}(V = v, W) \times r(O) \} \\ &= E_P(D_\eta \times r) \end{aligned}$$

In the first equality, M_P and M'_P are shorthand for the remaining terms in the expansion of the products. This concludes the proof of Proposition 2

Now, we derive the EIF for Ψ at $P \in \mathcal{M}$. From Proposition 1 and Proposition 2, after some simplifications, we conclude that the right-hand-side of eq. (17) can be written as

$$\begin{aligned} &\sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) \left\{ \left. \frac{d\eta(P_r(\alpha))(\mathbf{a}, v)}{d\alpha} \right|_{\alpha=0} - \left(\left. \frac{d\gamma(P_r(\alpha))(a_0, v)}{d\alpha} \right|_{\alpha=0} \right) m_{\Psi(P)}(\mathbf{a}, v) \right\} \\ &= \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) \{ E_P(D_\eta(P) \times r) - E_P(D_\gamma(P) \times r) m_{\Psi(P)}(\mathbf{a}, v) \} \\ &= E_P \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \tilde{h}(\mathbf{a}, v) [D_\eta(P) - D_\gamma(P) m_{\Psi(P)}(\mathbf{a}, v)] \times r(O) \right\} \\ &= E_P \{ D(Q, g, \Psi(P)) \times r \}, \end{aligned}$$

where

$$\begin{aligned}
 D(Q, g, \psi) &= \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \tilde{h}(\mathbf{a}, v) [D_\eta(P) - D_\gamma(P) m_{\Psi(P)}(\mathbf{a}, v)] \\
 &= \sum_{t=0}^K \sum_{\mathbf{a}} \tilde{h}(\mathbf{a}, V) C_t^{\mathbf{a}} \{Q_{t+1}^{\mathbf{a}}(\mathbf{L}_t, V, W) - Q_t^{\mathbf{a}}(\mathbf{L}_{t-1}, V, W)\} \\
 &\quad + \sum_{\mathbf{a}} \frac{I(A_0 = a_0)}{g(a_0 | W)} \left\{ \tilde{h}(\mathbf{a}, V) (Q_0^{\mathbf{a}}(V, W) - m_{\Psi(P)}(\mathbf{a}, V)) \right. \\
 &\quad \left. - E_V [\tilde{h}(\mathbf{a}, V) (Q_0^{\mathbf{a}}(V, W) - m_{\Psi(P)}(\mathbf{a}, V)) | A_0 = a_0, W] \right\} \\
 &\quad + \sum_{\mathbf{a}} E_V [\tilde{h}(\mathbf{a}, V) (Q_0^{\mathbf{a}}(V, W) - m_{\Psi(P)}(\mathbf{a}, V)) | A_0 = a_0, W] \\
 &\quad - \sum_{\mathbf{a}, v} \tilde{h}(\mathbf{a}, v) (\eta(P)(\mathbf{a}, v) - m_{\Psi(P)}(\mathbf{a}, v) \gamma(P)(a_0, v)).
 \end{aligned}$$

It follows from Remark 1 that the last summand $\sum_{\mathbf{a}, v} \tilde{h}(\mathbf{a}, v) (\eta(P)(\mathbf{a}, v) - m_{\Psi(P)}(\mathbf{a}, v) \gamma(P)(a_0, v)) = U(P, \Psi(P)) = 0$. To emphasize the role of P , we shall write $D(Q, g, \Psi(P))$ as $D(P)$. To see that $E_P D(P) = 0$, we first note that all but the last line are centered functions with respect to P ; secondly, from Remark 1, the last line also has zero expectation under P when $\psi = \Psi(P)$. Since we are operating under a saturated model, each component of $D(P)$ is in the tangent space, hence it is in fact the efficient influence curve.

To finish the proof of first part of Lemma 1, it suffices to see from eq. (17) that the normalizing quantity is given by the inverse of

$$\begin{aligned}
 \frac{\partial E_P D_W(P, \Psi(P))}{\partial \Psi(P)} &= \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} \gamma(P)(a_0, v) h(\mathbf{a}, v) \phi(\mathbf{a}, v) m_{\Psi(P)}(\mathbf{a}, v) [1 - m_{\Psi(P)}(\mathbf{a}, v)] \phi(\mathbf{a}, v)^\top \\
 &= \int_W P(W) \sum_{(\mathbf{a}, v)} P(v | A_0 = a_0, W) h(\mathbf{a}, v) \phi(\mathbf{a}, v) m_{\Psi(P)}(\mathbf{a}, v) [1 - m_{\Psi(P)}(\mathbf{a}, v)] \phi(\mathbf{a}, v)^\top \\
 &= E \left\{ \sum_{\mathbf{a}} \frac{I(A_0 = a_0)}{g(a_0 | W)} h(\mathbf{a}, V) \phi(\mathbf{a}, V) m_{\Psi(P)}(\mathbf{a}, V) [1 - m_{\Psi(P)}(\mathbf{a}, V)] \phi(\mathbf{a}, V)^\top \right\} \\
 &= M(\Psi(P), P).
 \end{aligned}$$

To prove the robustness statement, we first consider the case $Q(P) = Q(P_0)$. All but the last line in $D(P)$ are centered about components of $Q(P_0)$, therefore they will have expectation zero under P_0 . The last line in $E_{P_0} D(P)$ is $D_W(P_0, \Psi(P_0))$, which equals zero by definition of ψ_0 .

Next, we consider the case $g(P) = g_0$. By telescoping the sums in eq. (12) and applying definition of $Q_0^{\mathbf{a}}(P_0)$, we obtain

$$\begin{aligned}
 &P_0 D(Q(P), g_0, \psi_0) \\
 &= E_{P_0} \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} P_0(V = v | (A_0 = a_0, W)) \tilde{h}(\mathbf{a}, v) Q_0^{\mathbf{a}}(P_0)(V = v, W) \right\} \\
 &\quad - E_{P_0} \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} P_0(V = v | (A_0 = a_0, W)) \tilde{h}(\mathbf{a}, v) m_{\psi_0}(\mathbf{a}, v) \right\},
 \end{aligned}$$

which equals 0 by Remark 1.

D Proof of Lemma 2

To see that D^{A_0} is a score, up to a normalizing matrix, on the model where $g(A_0 | W)$ is known, we repeat the steps in Section C, but this time we only consider the class of 1-dimensional submodels indexed by

$$R_P = \{r = r' - E_P(r' | A_0, W) + E_P(r' | W) - E_P(r') : r' \in \mathcal{H}_P\} \subset \mathcal{H}_P.$$

It is straight forward to verify that $E_P(r | A_0 = a_0, W) = 0 = E_P(r | W)$ for such $r \in R_P$. Therefore, $D_{\gamma(P)(a_0, v)}$ in this case becomes $\frac{I(A_0 = a_0)}{g(a_0 | W)} I(V = v) - \gamma(P)(a_0, v)$ and

$$\begin{aligned}
 D_\eta(P) &= I(V = v) \sum_{t=0}^K C_t^{\mathbf{a}} \{Q_{t+1}^{\mathbf{a}}(\mathbf{L}_t, v, W) - Q_t^{\mathbf{a}}(\mathbf{L}_{t-1}, v, W)\} \\
 &\quad + \frac{I(A_0 = a_0)}{g(a_0 | W)} I(V = v) Q_0^{\mathbf{a}}(v, W) - \eta(\mathbf{a}, v).
 \end{aligned}$$

The rest is straightforward from eq. (16).

To see that $PD^{A_0}(Q, g, \Psi(P)) = 0$, it suffices to show that $PD_W^{A_0}(Q, g, \Psi(P)) = 0$:

$$\begin{aligned} & P \left\{ \frac{I(A_0 = a_0)}{g(a_0 | W)} \sum_{\mathbf{a}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) [Q_1^{\mathbf{a}}(L_0, V, W) - m_{\Psi(P)}(\mathbf{a}, v)] \right\} \\ &= P \left\{ \frac{I(A_0 = a_0)}{g(a_0 | W)} \sum_{\mathbf{a}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) \{Q_1^{\mathbf{a}}(L_0, V, W) - Q_0^{\mathbf{a}}(V, W)\} \right\} \\ &+ P \left\{ \frac{I(A_0 = a_0)}{g(a_0 | W)} \sum_{\mathbf{a}} h(\mathbf{a}, v) \phi(\mathbf{a}, v) [Q_0^{\mathbf{a}}(V, W) - m_{\Psi(P)}(\mathbf{a}, v)] \right\}. \end{aligned}$$

The first term in the right hand side of the equality is zero by definition of $Q_1^{\mathbf{a}}$; the second term is zero by Remark 1. There D^{A_0} is a valid estimating function.

To see its robustness properties under $g = g(P_0)$, first consider the case of $Q_t^{\mathbf{a}}(P) = Q_t^{\mathbf{a}}(P_0)$ for $t = 1, \dots, K$. Trivially, $P_0 D_t^*(Q, g) = 0$ for each $t \geq 1$ by definition of $Q_t^{\mathbf{a}}$. On the term $D_W^{A_0}$, we have

$$\begin{aligned} & P_0 \left\{ \frac{I(A_0 = a_0)}{g(P_0)(1 | W)} \sum_{\mathbf{a}} \tilde{h}(\mathbf{a}, v) [Q_1^{\mathbf{a}}(P_0)(L_0, V, W) - m_{\psi_0}(\mathbf{a}, v)] \right\} \\ &= P_0 \left\{ \sum_v Q^V(P_0)(v | A_0 = a_0, W) \sum_{\mathbf{a}} \tilde{h}(\mathbf{a}, v) [Q_0^{\mathbf{a}}(P_0)(V = v, W) - m_{\psi_0}(\mathbf{a}, v)] \right\} = 0, \end{aligned}$$

per definition of $Q_0^{\mathbf{a}}$ and ψ_0 .

Next, we consider the case of $g(P) = g_0$. By telescoping the sums in eq. (15) and applying definition of $Q_0^{\mathbf{a}}$, we again obtain

$$\begin{aligned} P_0 D^{A_0}(Q, g_0, \psi_0) &= P_0 \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} P_0(V = v | (A_0 = a_0, W)) \tilde{h}(\mathbf{a}, v) Q_0^{\mathbf{a}}(P_0)(V = v, W) \right\} \\ &- P_0 \left\{ \sum_{(\mathbf{a}, v) \in \mathcal{A} \times \mathcal{V}} P_0(V = v | (A_0 = a_0, W)) \tilde{h}(\mathbf{a}, v) m_{\psi_0}(\mathbf{a}, v) \right\} = 0. \end{aligned}$$

This completes proof of Lemma 2.

References

- [1] van der Laan MJ, Petersen M. Statistical learning of origin-specific statically optimal individualized treatment rules. *Int J Biostat.* 2007;3:Article 6.
- [2] van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. *Int J Biostat.* 2006;2:Article 11.
- [3] van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. Springer Series in Statistics, 1st ed. New York, NY: Springer, 2011.
- [4] Robins JM. Marginal structural models. In: Proceedings of the American Statistical Association. Section on Bayesian Statistical Science. 1997;1–10.
- [5] Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11:561–570.
- [6] Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Math Modell.* 1986;7:1393–1512.
- [7] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11:550–560.
- [8] van der Laan MJ, Petersen ML, Joffe MM. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *Int J Biostat.* 2005;1(1):Article 4–.
- [9] Petersen M, Deeks S, Martin J, van der Laan MJ. History-adjusted marginal structural models for estimating time-varying effect modification. *Am J Epidemiol.* 2007;166:985–993.
- [10] Petersen M, Deeks S, van der Laan MJ. Individualized treatment rules: generating candidate clinical trials. *Stat Med.* 2007;26:4578–601.
- [11] Petersen M, Porter K, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Technical report 269, Division of Biostatistics, University of California, Berkeley, 2010. <http://www.bepress.com/ucbbiostat/paper269>.
- [12] Pearl J. Causality: models, reasoning and inference, 2nd ed. New York: Cambridge University Press, 2009.
- [13] Robins JM, Rose S. Causal inference from complex longitudinal data. In: Berkane M, editor(s). *Latent Variable Modeling and Applications to Causality*. New York, NY: Springer Verlag, 1997:69–117.
- [14] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61:962–972.
- [15] van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Gene Molecul Biol.* 2007;6:Article 25.

- [16] Diaz I, van der Laan M. Doubly robust inference for targeted minimum loss-based estimation in randomized trials with missing outcome data. *Stat Med.* 2017;36:3807–3819.
- [17] Bickel PJ, Klaassen CA, Ritov Y, Wellner J. Efficient and adaptive estimation for semiparametric models. New York, NY: Springer-Verlag, 1997.
- [18] Ellis A, Dusetzina S, Hansen R, Gaynes B, Farley J, Stumer T. Investigating differences in treatment effect estimates between propensity score matching and weighting: a demonstration using STAR*D trial data. *Pharmacoepidemiol Drug Safety.* 2013;22:138–144.