Double Machine Learning for Sample Selection Models

Michela Bia*, Martin Huber**, and Lukáš Lafférs***

*Luxembourg Institute of Socio-Economic Research and University of Luxembourg, Esch-sur-Alzette, Luxembourg; **Department of Economics, University of Fribourg, Fribourg, Switzerland; ***Department of Mathematics, Matej Bel University, Banská Bystrica, Slovakia; ****Department of Economics, Norwegian School of Economics, Bergen, Norway

ABSTRACT
This article considers the evaluation of discretely distributed treatments when outcomes are only observed for a subpopulation due to sample selection or outcome attrition. For identification, we combine a selection-on-observables assumption for treatment assignment with either selection-on-observables or instrumental variable assumptions concerning the outcome attrition/sample selection process. We also consider dynamic confounding, meaning that covariates that jointly affect sample selection and the outcome may (at least partly) be influenced by the treatment. To control in a data-driven way for a potentially high dimensional set of pre- and/or post-treatment covariates, we adapt the double machine learning framework for treatment evaluation to sample selection problems. We make use of (a) Neyman-orthogonal, doubly robust, and efficient score functions, which imply the robustness of treatment effect estimation to moderate regularization biases in the machine learning-based estimation of the outcome, treatment, or sample selection models and (b) sample splitting (or cross-fitting) to prevent overfitting bias. We demonstrate that the proposed estimators are asymptotically normal and root-n consistent and investigate their finite sample properties in a simulation study. We also apply our proposed methodology to the Job Corps data. The estimator is available in the causalweight package for the statistical software R.

1. Introduction
In many studies aiming at evaluating the causal effect of a treatment or policy intervention, the empirical analysis is complicated by nonrandom outcome attrition or sample selection. Examples include the estimation of the returns to education when wages are only observed in the selective subpopulation of employed individuals or the effect of educational interventions like vouchers for private schools on college admissions tests when students non-randomly abstain from the test. Furthermore, in observational studies, treatment assignment is typically not random, implying that the researcher faces a double selection problem, namely selection into the treatment and observability of the outcome. A large literature addresses treatment selection by assuming a selection-on-observables assumption, implying that treatment is as good as randomly assigned conditional on observed pre-treatment covariates, see for instance the reviews by Imbens (2004) and Imbens and Wooldridge (2009). Furthermore, a growing number of studies addresses the question of how to control for the crucial confounders in a potentially high-dimensional vector of covariates in a data-driven way based on machine learning algorithms, see for instance the double machine learning framework of Chernozhukov et al. (2018).

In this article, we adapt the double machine learning framework to the evaluation of binary or multiply discrete treatments in the presence of sample selection or outcome attrition. In terms of identifying assumptions, we combine a selection-on-observables assumption for the treatment assignment with either selection-on-observables or instrumental variable assumptions concerning the outcome attrition/sample selection process. Such assumptions have previously been considered in Huber (2012) and Huber (2014b) for the estimation of the average treatment effect (ATE) based on inverse probability weighting, however, for pre-selected (or fixed) covariates. As methodological advancement, we derive doubly robust and efficient score functions for evaluating treatment effects under double selection and demonstrate that they satisfy so-called Neyman (1959) orthogonality. The latter property permits controlling for covariates in a data-driven way by machine learning-based estimation of the treatment, outcome, and attrition models under specific conditions. Therefore, the subset of important confounders need not be known a priori (but must be contained in the total set of covariates), which is particularly useful in high dimensional data with a vast number of covariates that could potentially serve as control variables.

We also consider dynamic confounding based on a sequential selection-on-observables assumption that is closely related to assumptions found in the dynamic treatment effect literature as for example in Robins (1986), Robins (1998), and Lechner (2009). This assumption permits that some covariates that jointly affect sample selection and the outcome may themselves be a function of the treatment as well as pre-treatment covariates, a scenario widely neglected in sample selection models.


despite its likely relevance in empirical applications. In particular when there is a substantial time lag between treatment assignment and the sample selection process, exploiting post-treatment covariates to tackle selection-outcome confounding seems more convincing than solely relying on pre-treatment covariates (as in conventional selection-on-observables assumptions) for addressing both treatment endogeneity and sample selection. For instance, such post-treatment covariates may include the value of the outcome and selection variables measured at a point in time after the treatment, but still prior to the outcome period of interest in which the effect on the outcome is measured and sample selection is to be controlled for. Consider for instance the evaluation of a training program on a wage outcome measured several years later, which is only observed conditional on employment. In this case, the post-treatment wage and employment histories prior to the outcome period of interest may serve as covariates to be controlled for, in addition to pre-treatment covariates (which may include pre-treatment employment and earnings). A crucial condition for the inclusion of post-treatment covariates is that they must not be affected by unobservables affecting the outcome after controlling for pre-treatment covariates and the treatment, otherwise controlling for post-treatment covariates entails endogeneity bias.

Following Chernozhukov et al. (2018), we show that treatment effect estimation based on our score functions (that are tailored to the various identifying assumptions) is root-\(n\) consistent and asymptotically normal under particular regularity conditions, in particular the \(n^{-1/4}\)-convergence of the machine learners. A further condition in the double machine learning framework is the prevention of overfitting bias due to correlations between the various estimation steps. This is obtained by estimating the treatment, outcome, and selection models on the one hand and the treatment effect on the other hand in different parts of the data. As in Chernozhukov et al. (2018), we subsequently swap the roles of the data parts and average over treatment effects in order to prevent asymptotic efficiency losses, a procedure known as cross-fitting. We also provide a simulation study (Appendix E) suggesting that our estimators perform decently in terms of the root mean squared error and coverage (by confidence intervals) in the simulation designs with several thousand observations considered. Finally, we present an empirical illustration considering the female sample of a study on Job Corps, a large training program for disadvantaged youth in the U.S. We apply our DML estimators to assess the effects of academic and vocational training on hourly wage, which is only observed conditional on employment, one and four years after program assignment and find some statistical evidence for positive longer-run impacts.

Our article is related to a range of studies tackling sample selection and selective outcome attrition. One strand of the literature models the attrition process based on a selection-on-observables assumption also known as missing at random (MAR) condition. The latter imposes conditional independence of sample selection and the outcome given observed information like the covariates and the treatment. Examples include Rubin (1976), Little and Rubin (1987), Carroll, Ruppert, and Stefanski (1995), Shah, Laird, and Schoenfeld (1997), Fitzgerald, Gottschalk, and Moffitt (1998), Abowd, Crepon, and Kramarz (2001), Wooldridge (2002), and Wooldridge (2007). Robins, Rotnitzky, and Zhao (1994), Robins, Rotnitzky, and Zhao (1995), and Bang and Robins (2005) discuss doubly robust estimators of the outcome that are consistent under MAR when either the conditional outcome or the attrition model are correctly specified. This approach satisfies Neyman orthogonality as required for double machine learning. However, their framework does not consider double selection into treatment and the observability of the outcome at the same time as we do in this article.

Negi (2020) suggests an alternative estimator under double selection that falls into the weighted \(M\)-estimation framework described in Sloczyński and Wooldridge (2018) and also satisfies doubly robustness, that is remains consistent under parametric misspecification of either the conditional outcome model or the treatment and selection models. This approach based on reweighting outcome models is nevertheless different to ours making use of efficient score functions and to the best of our knowledge, Neyman (1959) orthogonality (as required for double machine learning) has not been shown for weighted \(M\)-estimation (while we prove this property for our proposed estimators). A further difference is that Negi (2020) focuses on treatment evaluation when controlling for pre-treatment covariates to tackle double selection, while we in addition consider identification based on both pre- and post-treatment covariates (dynamic confounding) or an instrument for sample selection.

In contrast to MAR-based identification, so-called sample selection or nonignorable nonresponse models allow for unobserved confounders of the attrition process and the outcome. Unless strong functional form assumptions as in Heckman (1976), Heckman (1979), Hausman and Wise (1979), and Little (1995) are imposed, identification requires an instrumental variable (IV) for sample selection. We refer to Das, Newey, and Vella (2003), Newey (2007), Huber (2012), and Huber (2014b) for nonparametric estimation approaches in this context. To the best of our knowledge, this study is the first one to propose a doubly robust treatment effect estimator under nonignorable outcome attrition and to consider machine learning techniques to control for (possibly high-dimensional) covariates in this context. Our estimators are available in the causalweight package for R by Bodory and Huber (2018).

This article proceeds as follows. Using the potential outcome framework, Section 2 discusses the identification of the average treatment effect, either when outcomes are assumed to be missing at random (i.e., selection is on observables, as for the treatment) conditional on pre-treatment covariates or when outcome attrition is related to unobservables, known as nonignorable nonresponse, and an instrument is available for tackling this issue. Section 3 demonstrates identification under a sequential selection-on-observables which allows for dynamic confounding, meaning that outcomes are assumed to be missing at random conditional on pre- and post-treatment covariates. Section 4 proposes an estimator based on double machine learning and shows root-\(n\) consistency and asymptotic normality under

\(^1\) Relatively, Barnwell and Chaudhri (2020) consider several outcome periods under a monotonic MAR assumption (i.e., outcome attrition being an absorbing state weakly increasing over time) and also discuss the evaluation of randomly assigned treatments in this context based on the efficient score function. In contrast, our framework considers a single outcome period and permits selection into treatment to be related to observed confounders.
2. Identification Under Missingness at Random or Nonignorable Nonresponse

Our target parameter is the average treatment effect (ATE) of a binary or multiply discretely distributed treatment variable \( D \) on an outcome variable \( Y \). To define the effect of interest, we use the potential outcome framework, see Rubin (1974). Let \( Y(d) \) denote the potential outcome under hypothetical treatment assignment \( d \in \{0, 1, \ldots, Q\} \), with 0 indicating nontreatment and 1, \ldots, \( Q \) the different treatment choices (where \( Q \) is the number of nonzero treatments). The ATE when comparing two distinct treatment \( d \neq d' \) corresponds to \( \Delta = E[Y(d) - Y(d')] \) (where we omit \( d \) and \( d' \) in \( \Delta \) for the ease of notation). Furthermore, let \( Y \) denote the outcome realized under the treatment actually assigned to a subject, that is, \( Y = Y(D) \). Therefore, \( Y \) corresponds to the potential outcome under the treatment received, while the potential outcome under any counterfactual treatment remains unknown. A further complication is that treatment remains unknown. A further complication is that \( Y \) is assumed to be observed only for a subpopulation, that is, conditional on \( S = 1 \), where \( S \) is a binary variable indicating whether \( Y \) is observed/selected. Finally, we denote by \( X \) a set of observed covariates that are measured prior to treatment assignment. Throughout the article, we assume that the stable unit treatment value assumption (SUTVA, Rubin (1980)) holds such that \( \Pr(D = d \implies Y = Y(d)) = 1 \). This rules out interaction or general equilibrium effects and implies that the treatment is uniquely defined.

We subsequently invoke a set of assumptions previously considered in Huber (2012) and Negi (2020). The assumptions imply that both selection into the treatment and outcome attrition is related to observed variables only, such that the mean potential outcomes \( E[Y(d)] \) and the ATE \( \Delta \) are identified when controlling observables.

**Assumption 1 (Conditional Independence of the Treatment).** \( Y(d) \perp D|X = x \) for all \( d \in \{0, 1, \ldots, Q\} \) and \( x \) in the support of \( X \).

**Assumption 2 (Conditional Independence of Selection).** \( Y(d) \perp S|D = d', X = x \) for all \( d, d' \in \{0, 1, \ldots, Q\} \) and \( x \) in the support of \( X \).

**Assumption 3 (Common Support).** (a) \( \Pr(D = d|X = x) > 0 \) and (b) \( \Pr(S = 1|D = d, X = x) > 0 \) for all \( d \in \{0, 1, \ldots, Q\} \) and \( x \) in the support of \( X \).

By **Assumption 1**, no unobservables jointly affect the treatment and the potential outcomes conditional on covariates \( X \), while by **Assumption 2**, no unobservables jointly affect selection and the outcomes conditional on \( D, X \). By the so-called observational rule, stating that \( Y = Y(d) \) conditional on \( D = d \), **Assumption 2** implies that \( Y \perp S|D = d, X = x \), such that outcomes are missing at random (MAR) in the denomination of Rubin (1976). **Assumption 3(a)** is a common support restriction requiring that the conditional probability to receive a specific treatment given \( X \), henceforth referred to as treatment propensity score, is larger than zero in either treatment state. **Assumption 3(b)** requires that for any combination of \( D, X \), the conditional probability to be observed, henceforth referred to as selection propensity score, is larger than zero. Otherwise, the outcome is not observed for some combinations of these variables.

We henceforth denote by \( \mu(D, S, X) = E[Y|D, S, X] \) the conditional mean outcome, by \( p_d(X) = \Pr(D = d|X) \) and \( \pi(D, X) = \Pr(S = 1|D, X) \) the propensity scores, and by \( I[\cdot] \) the indicator function, which is equal to one if its argument is satisfied and zero otherwise. **Proposition 1** provides the identification of the mean potential outcomes (and thus, the ATE) based on the efficient score function.

**Proposition 1.** Under **Assumptions 1, 2, and 3**, the mean potential outcome in the total population is identified by

\[
E[Y(d)] = E[\psi_d], \quad \text{with} \quad \psi_d = I[D = d] \cdot S \cdot [Y - \mu(d, 1, X)] \frac{1}{p_d(X) \cdot \pi(d, X)} + \mu(d, 1, X)
\]

being the efficient score function.

**Proof of Proposition 1.** By taking the expectation of the first term in \( \psi_d \) and applying the law of iterated expectations as well as basic probability theory, we obtain

\[
E\left[ \frac{I[D = d] \cdot S \cdot [Y - \mu(d, 1, X)]}{p_d(X) \cdot \pi(d, X)} \right] = E\left[ \frac{E[I[D = d] \cdot S \cdot [Y - \mu(d, 1, X)]|X]}{p_d(X) \cdot \pi(d, X)} \right] = E[E[Y - \mu(d, 1, X)|D = d, S = 1, X]] = 0,
\]

given that **Assumption 3** holds such that \( p_d(X) \cdot \pi(d, X) > 0 \). Therefore, expression (1) is equivalent to the expectation of its second term:

\[
E[\mu(d, 1, X)] = E[E[Y(d)|D = d, S = 1, X]] = E[E[Y(d)|D = d, X]] = E[E[Y(d)|X]] = E[Y(d)],
\]

where the first equality follows from the observational rule \( Y = Y(d) \) conditional on \( D = d \), the second equality from **Assumption 2**, the third equality from **Assumption 1**, and the fourth equality from the law of iterated expectations. **Appendix C** follows Levy (2019) to formally show that \( \psi_d \) is the efficient score function.

In contrast to regression or weighting by the inverse of the propensity score as considered in Huber (2012), expression (1) is doubly robust in the sense that it identifies \( E[Y(d)] \) if either the conditional mean outcome \( \mu(d, 1, X) \) or the propensity scores \( p_d(X) \) and \( \pi(d, X) \) are correctly specified. Furthermore, it satisfies the so-called Neyman (1959) orthogonality, that is, is first-order insensitive to perturbations in \( \mu(D, S, X), p_d(X), \) and \( \pi(D, X), \) see **Appendix B.1**. This entails desirable robustness properties when using machine learning to estimate the outcome, treatment, and selection models in a data-driven way and is to the best of our knowledge a novel contribution to the
literature on treatment evaluation under outcome attrition. To
the best of our knowledge, Neyman orthogonality has not been
shown for the weighted M-estimator suggested by Negi (2020),
which is also doubly robust.

When sample selection or outcome attrition is related to
unobservables even conditional on observables, identification
generally requires an instrument for S. We therefore replace
Assumptions 2 and 3 by alternative ones that are closely related
to Huber (2014b), but maintain Assumption 1 (i.e., selection into
treatment is on observables).

**Assumption 4 (Instrument for Selection).** (a) There exists
an instrument Z that may be a function of D, that is, Z = Z(D),
is conditionally correlated with S, that is, E[Z · S|D, X] ≠ 0,
and satisfies (i) Y(d, z) = Y(d) and (ii) Y⊥Z|D = d, X = x
for all d ∈ {0, 1, . . . , Q} and x in the support of X,
(b) S = I{V ≤ χ(D, X, Z)}, where χ is a general function and V
is a scalar (index of) unobservable(s) with a strictly mono-
tonic cumulative distribution function conditional on X,
(c) V⊥(D, Z)|X.

Assumption 4 allows for an association of the unobservable V in
the selection equation and unobservables affecting the outcome,
such that Assumptions 1 and 2 generally do not hold conditional
on S = 1 due to the endogeneity of the post-treatment variable S.
Indeed, S = 1 implies that χ(D, X, Z) > V such that conditional
on X, the distribution of V generally differs across values of D.
This entails a violation of the conditional independence of D and
Y(d) given S = 1 and X if the potential outcome distributions
differ across values of V. We therefore require an instrumental
variable denoted by Z, which must not affect Y or be associated
with unobservables affecting Y conditional on D and X, as
invoked in 4(a).2 We apply a control function approach based
on this instrument,3 which requires further assumptions.

By the threshold crossing model postulated in 4(b), Pr(S =
1|D, X, Z) = Pr(V ≤ χ(D, X, Z)) = FV(χ(D, X, Z)), where
FV(ν) denotes the cumulative distribution function of V evalu-
ated at ν. We will henceforth use the notation Π = π(D, X, Z) =
Pr(S = 1|D, X, Z) for the sake of brevity. Again by Assumption
4(b), the selection probability Π increases strictly mono-
tonically in χ, such that there is a one-to-one mapping between
FV and values v given X. By Assumption 4(c), V is independent
of (D, Z) given X, implying that the distribution of V given X is
(nonparametrically) identified. By comparing individuals with
the same Π, we control for FV and thus, for confounding of
D and Y(d) by V conditional on S = 1, X. In other words,
Π serves as control function where the exogenous variation
comes from Z. Controlling for the distribution of V based
on the instrument is thus a feasible alternative to the (infeasible)
approach of directly controlling for V.

Figure 1 provides an illustration of a causal framework
that is in line with Assumptions 1 and 4, using a directed
acyclic graph with edges representing variables and arrows
representing causal effects, as for example, advocated in Pearl
(2000). U denotes unobservables affecting the outcome, which
may be arbitrarily associated with V, which affects selection.
Note that the dashed lines indicate that V, U are not observed.
Identification relies on instrument Z, which is not associated
with outcome Y conditional on D and X.

Furthermore, identification requires the following common
support assumption, which is similar to Assumption 3(a), but in
contrast to the latter also includes Π in the conditioning set.

**Assumption 5 (Common Support).** Pr(D = d|X = x, Π = π) >
0 for all d ∈ {0, 1, . . . , Q} and x, z in the support of X, Z.

In general, Assumption 5 requires the instrument Z to be
continuous and strong enough to importantly shift the selection
probability Π conditional on D, M, X in the selected pop-
ulation. Denoting by pd(X, Π) = Pr(D = d|X, Π) and
μ(D, S, X, Π) = E[Y|D, S, X, π(D, X, Z)], the following proposition
formally states our identification result based on the efficient
score function, which in contrast to previous approaches in
Huber (2012) and Huber (2014b) satisfies Neyman orthogo-
nality.4

**Proposition 2.** Under Assumptions 1, 4, and 5, the mean
potential outcome in the selected population is identified by

\[ E[Y(d)|S = 1] = E \left[ \phi_{d,S=1} | S = 1 \right], \]

\[ \phi_{d,S=1} = \frac{I(D = d) · [Y - \mu(d, 1, X, Π)]}{p_d(X, Π)} + \mu(d, 1, X, Π) \]

being the efficient score function.

The proof is provided in Appendix A.1.

We note that the selected population is a mixture of various
subpopulations that are defined in terms of how the treatment

---

2As an alternative set of IV restrictions, d’Haultfoeuille (2010) permits
the instrument to be associated with the outcome, but assumes conditional
independence of the instrument and selection given the outcome.

3See for example, Ahn and Powell (1993), Das, Newey, and Vella (2003), Newey
(2007), Newey, Powell, and Vella (1999), Blundell and Powell (2004), and
Imbens and Newey (2009) for further semi- and nonparametric control
function approaches in sample selection or instrumental variable models.

4While the efficient score function associated with (2) is technically speaking
doubly robust, that is, consistent if either μ(d, 1, X, Π) or pd(X, Π) is
correctly specified, it is worth noting that this property can generally only hold
if Π is correctly specified because it enters both μ(d, 1, X, Π) and pd(X, Π) as
first step estimator. However, our approach does not rely on (global) doubly
robustness but on Neyman orthogonality, which implies that DML is robust
to local perturbations in Π under particular regularity conditions.
affects selection into outcome observability, see for instance Frangakis and Rubin (2002), Zhang and Rubin (2003), and Zhang, Rubin, and Mealli (2008) for a more detailed discussion. When assuming a binary treatment, the selected population generally includes subpopulations that are either selected independently of treatment (always selected), or because of receiving treatment (but would not be selected without treatment), or because of not receiving the treatment (but would not be selected under treatment). For this reason, the ATE on the selected population is a weighted average of subpopulation-specific ATEs and may not always be the most interesting causal parameter. Lee (2009), for instance, focuses on the subpopulation of always selected, defined as those being employed ($S = 1$) independent of training participation (treatment $D$), when assessing the effect of training on wages. Under the additional assumption that the treatment does not decrease selection for anyone, reweighting selected observations based on the propensity of nontreatment ($p_0(X, \Pi)$) allows identifying the ATE among the always selected, as discussed in Section 2.4 of Huber (2014b). It follows that treatment propensity score-weighted version of expression (2) permits evaluating potential outcomes and causal effects within certain subpopulations, under the condition that the treatment never decreases selection.

Also the identification of the ATE in the total (rather than the selected) population requires further assumptions. The reason is that effects among selected observations cannot be extrapolated to the nonselected population if the effect of $D$ interacts with unobservables $U$ affecting the outcome, as $U$ is in general distributed differently across $S = 1, 0$ even conditional on $(X, \Pi)$ or $(D, X, \Pi)$. To see this, note that conditional on $\Pi = \text{Pr}(V \leq \chi(D, X, Z))$, the distribution of $V$ differs across the selected (satisfying $V \leq \chi(D, X, Z)$) and the nonselected (satisfying $V > \chi(D, X, Z)$), such that the distribution of $U$ differs, too, if $V$ and $U$ are associated. This generally implies that $E[Y(1) - Y(0)|S = 1, X, \Pi] \neq E[Y(1) - Y(0)|S = 0, X, \Pi]$. While control function $\Pi$ ensures (together with $X$) that the treatment is unconfounded in the selected subpopulation, it does not permit extrapolating effects to the nonselected population with unobserved outcomes, see also Huber and Melly (2015) for further discussion.

Assumption 6 therefore imposes homogeneity in the average treatment effect across selected and nonselected populations conditional on $X, V$. A sufficient condition for effect homogeneity is the separability of observed and unobserved components in the outcome equation, that is, $Y = \eta(D, X) + \nu(U)$, where $\eta, \nu$ are general functions and $U$ are unobserved terms. Furthermore, common support as postulated in Assumption 5 needs to be strengthened to hold in the entire population. In addition, the selection probability $\Pi$ must be larger than zero for any $d, x, z$ in their support. Otherwise, outcomes are not observed for some values of $D, X$. Assumption 7 formalizes this common support restriction.

5 Under separability, potential outcomes in the selected and nonselected populations might differ due to distinct levels of $U$, while conditional average treatment effects given $X, V$ are the same, because the influence of $U$ cancels out when taking differences in conditional outcomes across treatment states. See Huber (2014b) for further discussion.

### Assumption 6 (Conditional Effect Homogeneity). $E[Y(d) - Y(d')]|S = d, X = x, V = v] = E[Y(d) - Y(d')|X = x, V = v]$ for all $d \neq d' \in \{0, 1, \ldots, Q\}$ and $x, v$ in the support of $X, V$.

### Assumption 7 (Common Support). $\pi(d, x, z) > 0$ for all $d \in \{0, 1, \ldots, Q\}$ and $x, z$ in the support of $X, Z$.

### Proposition 3. Under Assumptions 1, 4, 5, 6, and 7, the ATE in the total population is identified by

$$
\Delta = E[\phi_d - \phi_{d'}], \quad \text{with} \quad \phi_d = \frac{E[D = d] \cdot \pi(d, X, Z)}{\rho_d(X, \Pi)} \cdot \pi(d, X, Z) + \mu(d, X, \Pi), \quad \text{being the efficient score function.}
$$

The proof is provided in Appendix A.2.

### 3. Identification Under Sequential Conditional Independence

In many applications, it might appear unrealistic that one can control for all variables jointly affecting the sample selection indicator by conditioning only on baseline covariates measured prior to treatment assignment, in particular when no instrument is at hand. This is particularly the case when there is a substantial time lag between treatment assignment and sample selection/attrition, which raises concerns about dynamic confounding. The latter implies that some confounders influencing both the outcome and sample selection are themselves a function of the treatment. We subsequently reconsider the MAR framework, but modify the identifying assumptions such that observed post-treatment confounders of $Y$ and $S$ are permitted. We will subsequently refer to observed post-treatment variables by $M$, in order to distinguish them from pre-treatment covariates $X$. For instance, if the outcome $Y$ and selection $S$ are assessed in a period denoted by $t$, then $M$ might be measured one period earlier in $t - 1$, while treatment $D$ is assigned still earlier in $t - 2$ and pre-treatment covariates $X$ are measured in $t - 3$. We note that $M$ may also include measures of $Y$ and $S$ in period $t - 1$. Identification is based on a sequential conditional independence, which is based on maintaining Assumption 1 (conditional independence of $D$ given $X$), but replacing Assumption 2 by a modified conditional independence assumption for the selection indicator $S$ that allows for dynamic confounding due to $M = M(D)$, that is, covariates possibly influenced by the treatment.

### Assumption 8 (Conditional Independence of Selection). $Y(d) \perp S|D = d', X = x, M = m$ for all $d, d' \in \{0, 1, \ldots, Q\}$ and $x, m$ in the support of $X$ and $M$.

By Assumption 8, no unobservables jointly affect selection and the potential outcome $Y(d)$ conditional on $D, X, M$, such that sample selection is selective with respect to observed characteristics only. When considering the nonparametric outcome and selection models $Y = \psi(D, X, M, U)$ and $S = \phi(D, X, M, V)$, Assumption 8 requires that unobservables $U$ and $V$ are independent. Assumption 8 differs from Assumption 2 in the sense that it allows for post-treatment variables that affect
both selection and the outcome in a later post-treatment period. For instance, the covariate pre-treatment health might not fully capture the health-induced effects on employment (S) and wages (Y). In this case, we should also control for post-treatment health (which might be affected by the treatment and may affect employment and wage in subsequent periods), as conditioning on pre-treatment health does not fully account for health-related sample selection. From this perspective, Assumption 8 appears more plausible than Assumption 2.

However, one important implication of Assumption 8 is that conditioning on post-treatment covariates M does not introduce endogeneity bias, meaning that conditional on D and X, there are no unobserved variables jointly affecting Y and M. This requires that all pre-treatment variables affecting both M and Y are included in the pre-treatment covariates X. For instance, if post-treatment health affects both selection (e.g., employment) and the outcome (e.g., wages) and is used as a control variable in M, then we must also control for pre-treatment health in X if pre-treatment health directly affects the outcome. Otherwise, post-treatment health is a bad control in the sense of Angrist and Pischke (2009), as conditioning on it introduces a spurious association between the outcome and the treatment via pre-treatment health.6

**Assumption 9 (Common Support).** (a) \( \Pr(D = d | X = x) > 0 \) and (b) \( \Pr(S = 1 | D = d, X = x, M = m) > 0 \) for all \( d \in \{0, 1, \ldots, Q\} \) and \( x, m \) in the support of \( X, M \).

Part (b) in Assumption 9 imposes a somewhat stronger common support restriction than part (b) in Assumption 3, as it requires the selection propensity score to be larger than zero for any combination of \( D, X, M \) (rather \( D, X \) only).

**Figure 2** provides an acyclic graph in which Assumptions 1 and 8 hold. Post-treatment covariates M may be influenced by \( D, X \) and might jointly affect \( S \) and \( Y \). Conditional on \( D, X, M \), there are, however, no unobservables jointly influencing \( S \) and \( Y \).

Let \( \mu(d, 1, X, M) = E[Y | D = d, S = 1, X, M] \) and \( \nu(d, 1, X) = E[E[Y | D = d, S = 1, X, M] | D = d, X] \) denote the conditional mean outcome and the nested conditional mean outcome, respectively.

**Proposition 4.** Under Assumptions 1, 8, 9, the mean potential outcome in the total population is identified by

\[
E(Y(d)) = E \left[ \theta_d \right], \quad \text{with} \quad \theta_d = \frac{1[D = d] \cdot [Y - \mu(d, 1, X, M)]}{\rho_d(X) \cdot \pi(d, 1, X, M)}
\]

The proof is provided in Appendix A.3.

We note that the efficient score function in expression (5) seems closely related to that for the evaluation of dynamic treatments, that is, sequential treatments that are assigned in different periods, see for example, Tran et al. (2019). Despite such similarities, the contexts differ, however, in terms of causal models and identification. In this article, we assess a single treatment while controlling for the partial observability of the outcome, whereas the literature on dynamic treatment effects evaluates alternative sequences of treatments, while typically not considering outcome attrition. To the best of our knowledge, we are the first to suggest such an efficient score function under sequential conditional independence in the context of outcome attrition.

### 4. Estimation by K-Fold Cross-Fitting

This section proposes DML estimators of the ATEs based on our various propositions, which we show to be root-n consistent under specific regularity conditions. We first suggest an estimation strategy for the mean potential outcome \( E[Y(d)] \) under MAR as discussed in Section 2 based on identification result (1). Let to this end \( W = \{W_i | 1 \leq i \leq n\} \) with \( W_i = (Y_i, S_i, D_i, X_i) \) for all \( i \) denote the set of observations in an iid sample of size \( n \). \( \eta \) denotes the plug-in (or nuisance) parameters, that is, the conditional mean outcome, treatment probability, and selection probability. Their respective estimates are referred to by \( \hat{\eta} = \{\hat{\mu}(D, 1, X), \hat{\rho}_d(X), \hat{\pi}(D, X)\} \) and the true parameters by \( \eta_0 = \{\mu_0(D, 1, X), \rho_{d0}(X), \pi_0(D, X)\} \). Finally, \( \psi_{d0} = E[\psi_d] \) denotes the true mean potential outcome under Proposition 1 as defined in (1).

We estimate \( \psi_{d0} \) by the following algorithm that combines the estimation of Neyman-orthogonal scores with sample splitting or cross-fitting and is root-n consistent under conditions outlined further below.

In order to obtain root-n consistency for the estimation of mean potential outcomes, we make the following assumption about the prediction qualities of machine learning for estimating the nuisance parameters. Following Chernozhukov et al. (2018),

---

6Related assumptions on the inclusion of post-treatment covariates for tackling confounding have been imposed in evaluations of dynamic (i.e., sequentially assigned) treatments, see Robbins (1986), Robins (1998), and Lechner (2009), or of causal mediation, see for example Imai and Yamamoto (2013) and Huber (2014a). Indeed, M is a mediator in the sense that part of the effect of D on Y operates via M. At the same time, M also affects selection S, which is the reason why it needs to be included as a control variable as stated in Assumption 8. One difference to the dynamic treatment effects literature, however, is that the selection indicator S does not affect (and is not affected by) the outcome, which is also a distinction with studies on surrogate outcomes (i.e., short-term outcomes through which the treatment effect on longer-term outcomes operates), see for instance Athey et al. (2019).
Algorithm 1 Estimation of $E[Y(d)]$ based on (1)

1. Split $\mathcal{W}$ in $K$ subsamples. For each subsample $k$, let $n_k$ denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}^c_k$ the complement set of all observations not in $k$.

2. For each $k$, use $\mathcal{W}^c_k$ to estimate the model parameters of the plug-ins $\mu(D,S=1,X,d), p_d(X), \pi(D,X)$ in order to predict these plug-ins in $\mathcal{W}_k$, where the predictions are denoted by $\hat{\mu}^k(D,1,X), \hat{p}_d^k(X), \hat{\pi}^k(D,X)$.

3. For each $k$, obtain an estimate of the score function (see $\psi_d$ in (1)) for each observation $i$ in $\mathcal{W}_k$, denoted by $\hat{\psi}_{d,i}$:

$$
\hat{\psi}^k_{d,i} = \frac{I(D_i = d) \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i)]}{\hat{p}_d^k(X_i) \cdot \hat{\pi}^k(d,X_i)} + \hat{\mu}^k(d,1,X_i).
$$

4. Average the estimated scores $\hat{\psi}^k_{d,i}$ over all observations across all $K$ subsamples to obtain an estimate of $\Psi_{d0}$ in the total sample, denoted by $\hat{\Psi}_d = 1/n \sum_{k=1}^K \sum_{i=1}^{n_k} \hat{\psi}^k_{d,i}$.



we introduce some further notation: let $(\delta_n)_{n=1}^\infty$ and $(\Delta_n)_{n=1}^\infty$ denote sequences of positive constants with $\lim_{n \to \infty} \delta_n = 0$ and $\lim_{n \to \infty} \Delta_n = 0$. Furthermore, let $c, \epsilon, C$ and $q$ be positive constants such that $q > 2$, and let $K \geq 2$ be a fixed integer. Also, for any random vector $R = (R_1, \ldots, R_l)$, let $\|R\|_q = \max_{1 \leq i \leq l} \|R_i\|_q$, where $\|R\|_q = \left( E[|R|^q] \right)^{\frac{1}{q}}$. In order to ease notation, we assume that $n/K$ is an integer. For the sake of brevity we omit the dependence of probability $Pr_{\mathcal{P}}$, expectation $E_{\mathcal{P}}(\cdot)$, and norm $\|\cdot\|_{p,q}$ on the probability measure $P$.

Assumption 10 (Regularity Conditions and Quality of Plug-in Parameter Estimates).

For all probability laws $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws the following conditions hold for the random vector $(Y,D,S,X)$ for $d \in \{0,1,\ldots,Q\}$:

(a) $\|Y\|_q \leq C$, $\|E[Y^2|D = d, S = 1, X]\|_\infty \leq C^2$,

(b) $Pr(\epsilon \leq p_d(X) \leq 1 - \epsilon) = 1$, $Pr(\epsilon \leq \pi_0(d,X)) = 1$,

(c) $\|Y - \mu_0(d,1,X)\|_2 = E[(Y - \mu_0(d,1,X))^2]^{\frac{1}{2}} \geq c$,

(d) Given a random subset $I$ of $[n]$ of size $n_k = n/K$, the nuisance parameter estimator $\hat{\mu}_I = \hat{\mu}_I(W_i\in\mathcal{W}^c_i)$ satisfies the following conditions. With $P$-probability no less than $1 - \Delta_n$:

$$
\|\hat{\mu} - \mu_0\|_q \leq C,
$$

$$
\|\hat{\mu} - 0\|_2 \leq \delta_n,
$$

$$
\|\hat{p}_d(X) - 1/2\|_{\infty} \leq 1/2 - \epsilon,
$$

$$
\|\hat{\pi}(D,X) - 1/2\|_{\infty} \leq 1/2 - \epsilon,
$$

$$
\|\hat{\mu}(D,S,X) - \mu_0(D,S,X)\|_2 \times \|\hat{p}_d(X) - p_d(X)\|_2 \leq \delta_n n^{-1/2},
$$

$$
\|\hat{\mu}(D,S,X) - \mu_0(D,S,X)\|_2 \times \|\hat{\pi}(D,X) - \pi_0(D,X)\|_2 \leq \delta_n n^{-1/2}.
$$

The only non-primitive condition is the condition (d), which puts restrictions on the quality of the nuisance parameter estimators. Condition (a) states that the distribution of the outcome does not have unbounded moments. (b) refines the common support condition such that the treatment and selection propensity scores are bounded away from 0 and 1, respectively. (c) states that covariates $X$ do not perfectly predict the conditional mean outcome.

For demonstrating the root-n consistency of our estimator of the mean potential outcome, we show that it satisfies the requirements of the DML framework in Chernozhukov et al. (2018) by first verifying linearity and Neyman orthogonality of the score (see Appendix B.1). As $\psi_d(W,\eta,\Psi_{d0})$ is smooth in $(\eta,\Psi_{d0})$, it then suffices that the plug-in estimators converge with rate $o(n^{-1/4})$ for achieving $n^{-1/2}$-convergence in the estimation of $\hat{\psi}_d$ see Theorem 1. A rate of $n^{-1/4}$ is attainable by many commonly used machine learners under specific conditions, such as lasso, random forests, boosting and neural nets, see for instance Belloni, Chernozhukov, and Hansen (2014), Kueck et al. (2023), Wager and Athey (2018), and Farrell, Liang, and Misra (2021).

To give an example, Belloni and Chernozhukov (2011) present a scenario under which linear lasso regression with Gaussian homoscedastic errors attains the desired convergence rate. Their Corollary 3.1 provides the following rate for lasso regression: $Op\left(\sqrt{\frac{\log(p)}{n}}\right)$, where $p$ is the total number of regressors, and $s$ is the number of a subset of regressors with nonzero coefficients. This implies that as long as $s^2 \log^2(p) = o(n)$, one attains the $o(n^{-1/4})$ rate for the nuisance parameter estimate, which holds under particular conditions and a specific data-driven choice of the lasso penalty. One crucial condition is approximate sparsity, which implies that the nuisance parameters can be represented as functions of $s \ll n$ regressors (whose identities are a priori unknown), up to a small approximation error. More concisely, the number of “important” regressors $s$ must be such that the approximation error is of the same order as the estimation error $\sqrt{s}/n$. We list a complete set of sufficient conditions in Appendix D.1.

Theorem 1. Under Assumptions 1–3 and 10, it holds for estimating $\Psi_{d0} = E[\psi_d]$ based on Algorithm 1:

$$
\sqrt{n}(\hat{\Psi}_d - \Psi_{d0}) \to N(0,\sigma_{\psi_d}^2), \quad \text{where} \quad \sigma_{\psi_d}^2 = E[(\psi_d - \Psi_{d0})^2].
$$

The proof is provided in Appendix B.1.

We subsequently discuss the estimation of $E[Y(d)]$ based on (4) and to this end define $\Phi_{d0} = E[\phi_d]$. We note that in this case, one needs to estimate the nested nuisance parameters $\mu(d,1,X,\Pi)$ and $p_d(X,\Pi)$, because they require the first-step estimation of $\Pi = \pi(D,X,Z)$. To avoid overfitting in the nested estimation procedure, the models for $\Pi$ on the one hand and $\mu(d,1,X,\Pi), p_d(X,\Pi)$ on the other hand are estimated in different subsamples. The plug-in estimates are now denoted by $\hat{\mu}_I = (\hat{\mu}(D,1,X,\Pi), \hat{p}_d(X,\Pi), \hat{\pi}(D,X,Z))$ and the true plug-ins by $\mu_0 = (\mu_0(D,1,X,\Pi), p_{d0}(X,\Pi), \pi_0(D,X,Z))$. Appendix D provides a formal discussion of sufficient conditions such that lasso estimation of the nested nuisance parameters attains specific convergence rates, which entail the satisfaction of certain regularity conditions (as stated in Assumption 11 in Appendix B.2.1, discussed below) required for the root-n-consistency of mean
potential outcome estimation. Specifically, we require that the second-step estimator of the nested conditional mean is smooth in small perturbations of the true value of the quantity estimated in the first step, which is the propensity score \( \Pi \). We provide support for this smoothness condition by drawing a connection between the lasso estimator and the quadratic programming literature.

**Algorithm 2** Estimation based on (4)

1. Split \( \mathcal{W} \) in \( K \) subsamples. For each subsample \( k \), let \( n_k \) denote its size, \( \mathcal{W}_k \) the set of observations in the sample and \( \mathcal{W}_k^c \) the complement set of all observations not in \( k \).
2. Split \( \mathcal{W}_k \) into 2 nonoverlapping subsamples and estimate the model parameters of \( \pi_0(D,X,Z) \) in one subsample and the model parameters of \( \mu_0(d,1,X,\Pi) \) and \( p_{d0}(X,\Pi) \) in the other subsample. Predict the plug-in models in \( \mathcal{W}_k \), where the predictions are denoted by \( \hat{\pi}^k(D,X,Z) \), \( \hat{p}^k_d(X,\hat{\Pi}^k) \), and \( \hat{\mu}^k(d,1,X,\hat{\Pi}^k) \).
3. For each \( k \), obtain an estimate of the efficient score function (see \( \phi_d \) in (4)) for each observation \( i \) in \( \mathcal{W}_k \), denoted by \( \hat{\phi}_{d,i}^k \):

\[
\hat{\phi}_{d,i}^k = \frac{I[D_i = d] \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)]}{\hat{p}^k_d(X_i,\hat{\Pi}^k) \cdot \hat{\pi}^k(d,1,X_i,Z_i)} + \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)
\]

4. Average the estimated scores \( \hat{\phi}_{d,i}^k \) over all observations across all \( K \) subsamples to obtain an estimate of \( \Phi_{d0} \) in the total sample, denoted by \( \hat{\Phi}_d = 1/n \sum_{k=1}^K \sum_{i=1}^{n_k} \hat{\phi}_{d,i}^k \).

An estimator of \( \Phi_{d0}^{S=1} = E[\phi_{d,S=1}|S = 1] \) based on (2) is obtained by two modifications in Algorithm 2. First, rather than relying on the total sample \( n \), one merely uses the subsample with observed outcomes which of size \( \sum_{i=1}^{n_k} S_i \) to split it into \( K \) subsamples. Second, in step 3, \( \hat{\phi}_{d,i}^k \) is to be replaced by

\[
\hat{\phi}_{d,S=1}^k = I[D_i = d] \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)]/\hat{p}^k_d(X_i,\hat{\Pi}^k) + \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)
\]

3. For each \( k \), obtain an estimate of the efficient score function (see \( \phi_d \) in (4)) for each observation \( i \) in \( \mathcal{W}_k \), denoted by \( \hat{\phi}_{d,i}^k \):

\[
\hat{\phi}_{d,i}^k = \frac{I[D_i = d] \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)]}{\hat{p}^k_d(X_i,\hat{\Pi}^k) \cdot \hat{\pi}^k(d,1,X_i,Z_i)} + \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)
\]

An estimator of \( \Phi_{d0}^{S=1} = E[\phi_{d,S=1}|S = 1] \) based on (2) is obtained by two modifications in Algorithm 2. First, rather than relying on the total sample \( n \), one merely uses the subsample with observed outcomes which of size \( \sum_{i=1}^{n_k} S_i \) to split it into \( K \) subsamples. Second, in step 3, \( \hat{\phi}_{d,i}^k \) is to be replaced by

\[
\hat{\phi}_{d,S=1}^k = I[D_i = d] \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)]/\hat{p}^k_d(X_i,\hat{\Pi}^k) + \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)
\]

An estimator of \( \Phi_{d0}^{S=1} = E[\phi_{d,S=1}|S = 1] \) based on (2) is obtained by two modifications in Algorithm 2. First, rather than relying on the total sample \( n \), one merely uses the subsample with observed outcomes which of size \( \sum_{i=1}^{n_k} S_i \) to split it into \( K \) subsamples. Second, in step 3, \( \hat{\phi}_{d,i}^k \) is to be replaced by

\[
\hat{\phi}_{d,S=1}^k = I[D_i = d] \cdot S_i \cdot [Y_i - \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)]/\hat{p}^k_d(X_i,\hat{\Pi}^k) + \hat{\mu}^k(d,1,X_i,\hat{\Pi}^k)
\]

**Theorem 3.** Under Assumptions 1, 4, 6, 7, and 11, it holds for estimating \( \Phi_{d0} = E[\phi_d] \) based on Algorithm 2: \( \sqrt{n}(\hat{\Phi}_d - \Phi_{d0}) \rightarrow N(0,\sigma_{\Phi_d}^2) \), where \( \sigma_{\Phi_d}^2 = E[(\phi_d - \Phi_{d0})^2] \).

The proofs are provided in Appendices B.2 and B.3.

Next, we consider the estimation of \( \theta_{d0} = E[\theta_d] \) based on (5). Similarly to estimation based on (4), this requires us to estimate a nested nuisance parameter, namely \( \nu(d,1,X) = E(\nu(d,1,X,M)|D = d,X) \). To avoid overfitting in the nested estimation procedure, the models for \( \mu(d,1,X,M) \) and \( \nu(d,1,X) \) are estimated in different subsamples. We refer to Appendix D for primitive conditions that are sufficient for the nested lasso estimator to attain specific convergence rates which satisfy certain regularity conditions (see Assumption 12) that are required for the root-n-consistency of mean potential outcome estimation. Under the regularity conditions stated in Assumption 12 in Appendix B.4.1, which are similar to those stated earlier but include one additional rate condition, and the sequential conditional independence assumption, estimation based on Algorithm 3 (which is a trivial modification of Algorithm 2 and is stated in Appendix B.4.2) is asymptotically normal, as postulated in Theorem 4.

**Theorem 4.** Under Assumptions 1, 8, 9, and 12, it holds for estimating \( \Theta_{d0} = E[\theta_d] \) based on Algorithm 3:

\[
\sqrt{n}(\hat{\Theta}_d - \Theta_{d0}) \rightarrow N(0,\sigma_{\Theta_d}^2), \text{ where } \sigma_{\Theta_d}^2 = E[(\theta_d - \Theta_{d0})^2].
\]

The proof of Theorem 4 is provided in Appendix B.4.

**5. Application**

As an empirical illustration, we apply our method to the Job Corps (JC) training program. The data come from the National Job Corps Study (NJCS), a randomized social experiment conducted in the mid-to-late 1990s in the United States to evaluate the effectiveness of JC on different labor market outcomes. The JC is the largest and most comprehensive job training program for disadvantaged youth in the United States, in which participants are exposed to different types of academic and vocational instruction. The dataset contains very detailed pre-treatment information about program participants, such as: expectations, motivations for applying to JC, age, gender, number of children at the moment of treatment assignment, occupation, household income, hourly wage, educational level, marital status, whether the individual was previously attending a school, JC program, or some other academic or vocational training, health status, past employment, types of crimes committed, family support for attending the training, and information on the mother and the father (e.g., education and employment). Furthermore, a large range of variables for instance related to labor market status, employment, income, and education are reassessed in several follow-up interviews after JC assignment.

Schochet, Burghardt, and Glazerman (2001) and Schochet, Burghardt, and McConnell (2008) evaluate the impact of random program assignment on a wide range of labor market outcomes, showing positive effects on education, employment, and
earnings in the longer run. Several studies focus on more specific program aspects of JC, as the effect of the length of exposure to training or of specific training sequences on labor market outcomes like employment and earnings, see for example, Flores et al. (2012) and Bodory, Huber, and Laffère (2022). Other contributions assess JC’s causal mechanisms, that is, the program’s direct and indirect effects (operating via specific mediating variables) on labor market outcomes and health, see for example, Flores and Flores-Lagunes (2009), Flores and Flores-Lagunes (2010), Huber (2014a), and Frölich and Huber (2017).

Some studies aim at tackling the sample selection issue related to wages, which are only observed conditional on employment. For instance, Frumento et al. (2012) and Zhang, Rubin, and Mealli (2009) use the principal stratification approach of Frangakis and Rubin (2002) to evaluate the effects of JC on employment and wages among specific subgroups, for example those finding employment irrespective of training participation, rather than among the full population. Frumento et al. (2012) simultaneously address the several identification issues related to noncompliance of training participation with JC assignment as well as missingness in wage outcomes due to survey nonresponse or nonemployment based on a maximum likelihood approach using finite mixture models. Lee (2009) applies a partial identification approach for bounding the effect of JC assignment on wage among those finding employment irrespective of the assignment. Rather than invoking MAR or IV assumptions for sample selection, this method merely assumes that randomization into JC never decreases employment, at the cost of giving up point identification. Semenova (2020) suggests a DML approach to tighten the bounds by controlling for covariates X in a data-driven way.

For our empirical analysis we similarly to Frölich and Huber (2017) focus on female applicants to JC, among whom the employment share is generally lower than among male applicants. We aim at estimating the effects of academic or vocational training received in the first year of the program (D) on hourly wage (Y) in the short run measured in the last week of the first year or in the longer run, measured 4 years after random assignment (RA) to JC. Hourly wage is only observed conditional on employment (S) in the respective outcome period. Even though JC assignment is random, actual participation in training activities is likely selective and associated with individual factors, similarly to the selection into employment. As for example discussed in Lechner and Wunsch (2013) and Biewen et al. (2014), the previous labor market history and socio-economic characteristics are likely important confounders when assessing the impact of training interventions, which motivates our DML approach to account for a rich set of covariates in a data-driven way. To assess the short-run effects, we assume MAR as discussed in Section 2 and use our DML approach based on Theorem 1 to control for our all in all 355 baseline covariates (X). We do not consider the IV-based approach underlying Theorems 2 and 3, due to the lack of a plausible instrument that is of rich support and sufficiently strong in influencing employment conditional on the covariates.

For assessing the longer run effects, we invoke the sequential conditional independence assumption of Section 3 to apply DML based on Theorem 4. To this end, we additionally control for 619 and 156 post-treatment covariates (M) in the second and third year after JC assignment, respectively. For instance, we condition on detailed information about post-treatment labor market behavior, namely hours worked, employment, and earnings at different points in time of the second and third years after program start, that is, prior to assessing the effect on the wage outcome four years after program assignment. Assumption 8 imposes that conditional on these post-treatment covariates and outcome values that are measured up to one period prior to the outcome period of interest, as well as pre-treatment covariates measured prior to JC start and program participation, sample selection (employment) in the outcome period (four years after assignment) is independent of wages in the outcome period.

Appendix F presents descriptive statistics for selected variables in X and M. We also refer to Bodory, Huber, and Laffère (2022) for a more detailed description of the pre- and post-treatment covariates used in our application and note that all numeric variables have been standardized to have a mean equal to 0 and standard deviation equal to 0.5 to facilitate the machine learning-based estimation of the nuisance parameters. For estimation, we apply the treatselDML and dyn treatDML commands of the causalweight package for R, using 3-fold cross-fitting and the random forest (with default options of the SuperLearner package) as machine learner. The latter is a nonparametric approach allowing for nonlinear associations between the outcome, treatment, and selection on the one hand and the covariates on the other hand.

Table 1 reports the total number of females randomized into JC for whom participation in either vocational (843) or academic training (830) in the first year after program assignment is registered in the data. 200 females did not participate in any JC training activity in the first year and serve as the control group in our analysis. Furthermore, 1698 were randomized out of JC and are not considered in our analysis. Table 2 provides information on the average outcomes, namely the mean hourly wage by training status measured 1 and 4 years after assignment.

Table 1. Treatment distribution.

| treatment                  | Observations |
|----------------------------|--------------|
| Randomized out of JC       | 1698         |
| Control group (no training)| 200          |
| Academic training          | 830          |
| Vocational training        | 843          |

NOTE: column “treatment” provides the treatment status of females: randomized out of JC (not part of our analysis), control group (randomized in but not participating in any program in the first year), participating in academic training in the first year, or participating in vocational training in the first year. Column “observations” shows the number of observations by treatment status.

Table 2. Descriptive statistics of the hourly wage 1 and 4 years after RA.

|                          | Any training | No training | Academic | Vocational |
|--------------------------|--------------|-------------|----------|------------|
| Mean wage 1 year after assignment | 1.845        | 1.850       | 1.625    | 2.063      |
|                          | (2.965)      | (3.150)     | (2.872)  | (3.040)    |
| Mean wage 4 years after assignment | 5.507        | 5.210       | 5.044    | 5.962      |
|                          | (3.558)      | (3.449)     | (3.612)  | (3.445)    |

NOTE: columns “Any training”, “No training”, “Academic”, and “Vocational” provide the mean hourly wage among females taking any form of (academic or vocational) training in the first year, females in the control group (no training), females in academic training, and females in vocational training, respectively. Standard deviations are reported in parentheses.
to JC, respectively, as well as the respective standard deviations, which are reported in parentheses.

Table 3 reports the ATE estimates for academic and vocational training based on our DML approaches. The upper panel provides the short-run effects on hourly wages in the last week of the first year when assuming MAR. The point estimates of both academic and vocational training are positive (0.270 and 0.189 U.S. $, respectively), but neither effect is statistically significant at any conventional level. The lower panel provides the longer-run effects on hourly wages 4 years after assignment based on the sequential conditional independence assumption. Both ATE estimates are again positive, but only the effect of vocational training, which amounts to an hourly increase of 0.567 $, is statistically significant (at the 1% level). Our findings therefore suggest that JC-based education may foster human capital accumulation in a way that increases hourly wages after several years, in particular through vocational training. In contrast, they do not provide a clear conclusion concerning the short-run effects, which are positive, but less precisely estimated than the longer-run effects.

6. Conclusion

In this article, we discussed the evaluation of average treatment effects in the presence of sample selection or outcome attrition based on double machine learning. In terms of identifying assumptions, we imposed a selection-on-observables assumption on treatment assignment, which was combined with either selection-on-observables or instrumental variable assumptions concerning the outcome attrition/sample selection process. We also considered a sequential selection-on-observables assumption allowing for dynamic confounding such that covariates jointly affecting the outcome and sample selection may be affected by the treatment, which avoids exclusively relying on pre-treatment covariates. We proposed doubly robust score functions and formally showed the satisfaction of Neyman orthogonality, implying that estimators based on these score functions are robust to moderate (local) regularization biases in the machine learning-based estimation of the outcome, treatment, or sample selection models. Furthermore, we demonstrated the root-n consistency and asymptotic normality of our double machine learning approach to average treatment effect estimation under specific regularity conditions. We also provided an empirical illustration to the US Job Corps data, in which we assessed the effects of training on hourly wage one and four years after program assignment and found some statistical evidence for positive longer-run impacts. Our estimation procedure is available in the causalweight package for the statistical software R.

Supplementary Materials

The appendices include the following: proofs of propositions (A), proofs of theorems (B), derivation of influence functions (C), a discussion on the convergence rates of ML estimators (D), a simulation study (E), and descriptive statistics related to the empirical example (F). We also provide a replication package for the empirical example in the main paper.

Acknowledgments

We have benefited from comments by Alyssa Carlson, David Kaplan, Jannis Kueck, Peter Mueser, and seminar participants at the University of Missouri.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

Laffers acknowledges support provided by the Slovak Research and Development Agency under contract no. APVV-21-0360 and VEGA-1/0398/23. Michela Bia acknowledges financial support from the Inter Mobility IN Program: “Causal Mediation Analysis and Machine Learning based estimators (CAME),” funded by the Luxembourg National Research Fund.

ORCID

Michela Bia http://orcid.org/0000-0002-0892-7973
Martin Huber http://orcid.org/0000-0002-8590-9402
Lukáš Laffers http://orcid.org/0000-0002-3141-3591

References

Abowd, J., Crepon, B., and Kramarz, F. (2001), “Moment Estimation With Attrition: An Application to Economic Models,” Journal of the American Statistical Association, 96, 1223–1230. [959]
Ahn, H., and Powell, J. (1993), “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism,” Journal of Econometrics, 58, 3–29. [961]
Angrist, J. D., and Pischke, J.-S. (2009), Mostly Harmless Econometrics: An Epiricist’s Companion, Princeton: Princeton University Press. [963]
Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019), “The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely,” Discussion paper, National Bureau of Economic Research. [963]
Bang, H., and Robins, J. (2005), “Doubly Robust Estimation in Missing Data and Causal Inference Models,” Biometrika, 61, 962–972. [959]
Barnwell, J.-L., and Chaudhuri, S. (2020), “Efficient Estimation in Sub and Full Populations with Monotonically Missing at Random Data,” working paper, McGill University, Montreal. [959]
Belloni, A., and Chernozhukov, V. (2011), High Dimensional Sparse Econometric Models: An Introduction, Berlin: Springer. [964]
Belloni, A., Chernozhukov, V., and Hansen, C. (2014), “Inference on Treatment Effects After Selection among High-Dimensional Controls,” The Review of Economic Studies, 81, 608–650. [964]
Biewen, M., Fitzeenberger, B., Osikominu, A., and Paul, M. (2014), “The Effectiveness of Public-Sponsored Training Revisited: The Importance of Data and Methodological Choices,” Journal of Labor Economics, 32, 837–897. [966]
Blundell, R. W., and Powell, J. L. (2004), “Endogeneity in Semiparametric Binary Response Models,” The Review of Economic Studies, 71, 655–679. [961]
Bodory, H., and Huber, M. (2018), “The Causalweight Package for Causal Inference in R,” SES working paper 493, University of Fribourg. [959]
Tran, L., Yiannoutsos, C., Wools-Kaloustian, K., Siika, A., van der Laan, M., and Petersen, M. (2019), “Double Robust Efficient Estimators of Longitudinal Treatment Effects: Comparative Performance in Simulations and a Case Study,” *The International Journal of Biostatistics*, 15, 1–27. [963]

Wager, S., and Athey, S. (2018), “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” *Journal of the American Statistical Association*, 113, 1228–1242. [964]

Wooldridge, J. (2002), “Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification,” *Portuguese Economic Journal*, 1, 141–162. [959]

——— (2007), “Inverse Probability Weighted Estimation for General Missing Data Problems,” *Journal of Econometrics*, 141, 1281–1301. [959]

Zhang, J., Rubin, D., and Mealli, F. (2009), “Likelihood-based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification,” *Journal of the American Statistical Association*, 104, 166–176. [966]

Zhang, J., and Rubin, D. B. (2003), “Estimation of Causal Effects via Principal Stratification When Some Outcome are Truncated by Death,” *Journal of Educational and Behavioral Statistics*, 28, 353–368. [962]

Zhang, J., Rubin, D. B., and Mealli, F. (2008), “Evaluating the Effects of Job Training Programs on Wages through Principal Stratification,” in *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics* (Vol. 21), eds. D. Millimet, J. Smith, and E. Vructil, pp. 117–145, Bingley: Emerald Group Publishing Limited. [962]