Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials

Nima S. Hejazi\(^1,2\) | Mark J. van der Laan\(^3,4\) | Holly E. Janes\(^5,6,7\) | Peter B. Gilbert\(^5,6,7\) | David C. Benkeser\(^8\)

\(^1\) Graduate Group in Biostatistics, University of California, Berkeley, California
\(^2\) Center for Computational Biology, University of California, Berkeley, California
\(^3\) Division of Epidemiology & Biostatistics, School of Public Health, University of California, Berkeley, California
\(^4\) Department of Statistics, University of California, Berkeley, California
\(^5\) Vaccine & Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington
\(^6\) Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington
\(^7\) Department of Biostatistics, University of Washington, Seattle, Washington
\(^8\) Department of Biostatistics & Computational Biology, Rollins School of Public Health, Emory University, Atlanta, Georgia

**Correspondence**
Nima S. Hejazi, Graduate Group in Biostatistics, Center for Computational Biology, University of California, Berkeley, CA 94720.
Email: nhejazi@berkeley.edu

**Funding information**
U.S. National Library of Medicine, Grant/Award Number: T32LM012417; National Institute Of Allergy And Infectious Diseases of the National Institutes of Health, Grant/Award Number: R01AI074345

**Abstract**
The advent and subsequent widespread availability of preventive vaccines has altered the course of public health over the past century. Despite this success, effective vaccines to prevent many high-burden diseases, including human immunodeficiency virus (HIV), have been slow to develop. Vaccine development can be aided by the identification of immune response markers that serve as effective surrogates for clinically significant infection or disease endpoints. However, measuring immune response marker activity is often costly, which has motivated the usage of two-phase sampling for immune response evaluation in clinical trials of preventive vaccines. In such trials, the measurement of immunological markers is performed on a subset of trial participants, where enrollment in this second phase is potentially contingent on the observed study outcome and other participant-level information. We propose nonparametric methodology for efficiently estimating a counterfactual parameter that quantifies the impact of a given immune response marker on the subsequent probability of infection. Along the way, we fill in theoretical gaps pertaining to the asymptotic behavior of nonparametric efficient estimators in the context of two-phase sampling, including a multiple robustness property enjoyed by our estimators. Techniques for constructing confidence intervals and hypothesis tests are presented, and an open source software implementation of the methodology, the \texttt{txshift} R package, is introduced. We illustrate the proposed techniques using data from a recent preventive HIV vaccine efficacy trial.

**Keywords**
causal inference, stochastic interventions, targeted minimum loss estimation, two-phase sampling, vaccine efficacy
1 | INTRODUCTION

Ascertaining the population-level causal effects of exposures is a common goal in scientific research. Such effects can be formulated via summaries of the distribution of counterfactual random variables, which describe the values a measurement would have taken if a particular level of exposure were assigned to the unit. Often, the exposure of interest is continuous-valued—for example, the dose of a drug or level of an immune response marker induced by a vaccine. We consider the latter in the context of a phase IIb trial of a vaccine to prevent infection by human immunodeficiency virus (HIV), the HIV Vaccine Trials Network’s (HVTN) 505 efficacy trial (Hammer et al., 2013). A key secondary objective of the trial was to evaluate the role of vaccine-induced immune responses in generating protective efficacy against HIV infection (Janes et al., 2017). Identification of immune response markers causally related to protection is critical both for understanding of the biological mechanisms of a vaccine and for guiding the development of future vaccines.

To study such relationships, it is natural to consider a dose-response curve that summarizes vaccinated participants’ risk of HIV infection as a function of a particular immune response marker. A causal formulation of such a dose-response analysis would consider a (possibly infinite) collection of counterfactual outcomes, each representing the HIV infection risk that would have been observed if all individuals’ immune responses had been set to a particular level. Studying how the proportion of infected individuals varies as a function of the level of an immune response marker could provide insights into causal mechanisms underlying the vaccine’s effects. Unfortunately, several difficulties arise when considering such a dose-response approach. Nonparametric estimation and inference on the causal dose-response curve is challenging and requires nonstandard techniques (e.g., Kennedy et al., 2017). More importantly, this approach may require consideration of counterfactual variables that are scientifically unrealistic. Namely, it may be impossible to imagine a world where every vaccinated participant exhibits high immune responses, simply due to phenotypic variability of participants’ immune systems. This calls into question the validity of counterfactual dose-response analysis strategies that evaluate the effects of immune response markers.

An alternative framework for assessing effects of continuous-valued exposures involves counterfactual outcomes resulting from stochastic interventions (Díaz and van der Laan, 2012; Haneuse and Rotnitzky, 2013). Although static interventions set the same level of an exposure to all units, stochastic interventions instead set exposure level equal to a random draw from a particular distribution. This provides a more flexible approach for defining counterfactual random variables. Indeed, static interventions are a special case of stochastic interventions in which the intervention mechanism is drawn from a degenerate distribution with point mass on a single value. To define meaningful counterfactuals, care must be taken in defining the distribution from which the exposure is drawn. One strategy is to draw from a modified version of the true exposure distribution—the natural distribution of exposure under no intervention. For example, one may consider drawing immune response markers from a distribution similar to the naturally observed post-vaccination distribution of immune response markers but that has been shifted upward (or downward) for some or all participants. Counterfactuals defined by such interventions may be better aligned with plausible future interventions, such as refinements of the current vaccine that provide improved immune responses. Evaluating the population-level risk of HIV infection under such interventions is useful for several reasons. First, this measure of risk provides a relevant mechanism by which to rank-order immune response markers by their importance for HIV infection risk. Such information could be used in a future phase 2a trial of a refined candidate HIV vaccine to define go/no-go criteria based on immune response marker endpoints for advancing the vaccine to an efficacy trial. Relatedly, the risk measure may be useful for “transport formulas” that predict vaccine efficacy in new settings different from that in the original efficacy trial (Pearl and Bareinboim, 2014).

The above highlights the need for methodology to identify and estimate population-level causal effects of stochastic interventions; recent work has provided several candidate approaches (Díaz and van der Laan, 2012; Haneuse and Rotnitzky, 2013). These works show conditions for the identification and detail several estimators of parameters of distributions of counterfactuals defined by stochastic interventions. However, these approaches are not directly applicable to studies like HVTN 505 trial, where a two-phase, case-cohort sampling design was used to measure participants’ immune responses. Under this design, all vaccine recipients diagnosed with HIV-1 infection (i.e., “cases”) had their post-vaccination immune responses measured, while only a random sample of HIV-uninfected vaccine recipients had immune responses measured (Janes et al., 2017). This sampling design complicates the estimation of causal effects, as the cohort definition depends on post-randomization data, namely whether a participant was infected. Rose and van der Laan (2011), among others, discuss strategies for efficient estimation under two-phase sampling designs, emphasizing an inverse probability of censoring weighted modification that may be coupled with targeted minimum loss estimation to account for study design. Their approach yields an asymptotically linear estimator so long as the probability...
of inclusion in the second-phase sample is known by design or estimable via maximum likelihood. This latter requirement precludes usage of their proposed estimators in situations where, for example, sampling probabilities are unknown and may depend on continuous-valued covariates. Although the term “two-phase sampling” has traditionally been used to refer to outcome-dependent Bernoulli or without replacement sampling based on discrete covariates, recent efforts have extended the concept to the usage of continuous-valued covariates in constructing second-phase samples (e.g., Gilbert et al., 2014). Rose and van der Laan (2011) suggested a more complicated procedure that could be appropriate in such settings, but it has neither been evaluated in simulation nor data analysis.

In the present work, we develop estimators of the mean counterfactual outcome under a stochastic intervention when the exposure is measured via two-phase sampling. We provide several contributions to literatures on two-phase sampling and stochastic interventions. To the former, we (a) formalize assumptions needed for efficient nonparametric inference under two-phase sampling; (b) characterize a multiple robustness of the estimators; and (c) provide the first comparison of the practical performance of these estimators. Our contributions to the literature on stochastic interventions are (a) a novel estimator of a conditional density that is valid under two-phase sampling, while achieving a fast convergence rate, a crucial development for generating efficient estimators of the mean counterfactual; and (b) an extension of nonparametric inference on mean counterfactuals under stochastic interventions using projections onto nonparametric working marginal structural models. Finally, we provide open source R packages, ttxshift (Hejazi and Benkeser, 2020a, 2020b) and haldensify (Hejazi et al., 2020) that implement our proposed estimators.

2 | PRELIMINARIES AND BACKGROUND

2.1 | Notation, data, and target parameter

Consider data generated by typical cohort sampling represented by the random variable \(X = (W, A, Y)\), where \(W \in \mathcal{W}\) is a vector of baseline covariates, \(A \in \mathcal{A}\) a real-valued exposure, and \(Y \in \mathcal{Y}\) an outcome of interest. Initially, we assume access to \(n\) independent copies of \(X\), using \(P_0^X\) to denote the distribution of \(X\). We assume a nonparametric statistical model \(M^X\) for \(P_0^X\). We denote by \(q_{0,Y}\) the conditional density of \(Y\) given \(\{A, W\}\) with respect to some dominating measure, \(q_{0,A}\) the conditional density of \(A\) given \(W\) with respect to dominating measure \(\mu\), and \(q_{0,W}\) the density of \(W\) with respect to dominating measure \(\nu\). We use \(p_0^X\) to denote the density of \(X\) with respect to the product measure. This density evaluated on a typical observation \(x\) may be expressed \(p_0^X(x) = q_{0,Y}(y \mid A = a, W = w)q_{0,A}(a \mid W = w)q_{0,W}(w)\).

To define a counterfactual quantity of interest, we introduce a nonparametric structural equation model (NPSEM) (Pearl, 2000), which assumes \(X\) is generated by the following system of structural equations: \(W = f_W(U_W); A = f_A(W, U_A); Y = f_Y(A, W, U_Y)\). Here, \(f\) are deterministic functions and \(\{U_W, U_A, U_Y\}\) are exogenous random variables such that \(U_A \perp U_Y\) and either \(U_W \perp U_Y\) or \(U_W \perp U_A\), which establishes that conditioning on \(W\) is sufficient to control confounding of \(A\) and \(Y\).

The NPSEM parameterizes \(p_0^X\) in terms of the distribution of the random variables \((X, U)\) and implies a model for the distribution of counterfactual random variables generated by interventions on the data-generating process. For example, a static intervention replaces \(f_A\) with a real number \(a\). A stochastic intervention replaces the value \(A\) would naturally assume with a draw from a post-intervention distribution \(q_{0,A}(\cdot \mid W)\), where the zero subscript is included to emphasize that \(q_{0,A}\) may depend on \(P_0^X\). A static intervention may be viewed as a stochastic intervention where \(q_{0,A}(\cdot \mid W)\) places all mass on a single point. Diaz and van der Laan (2012) described a stochastic intervention that draws \(A\) from a distribution such that for a real number \(a\), \(\tilde{q}_{0,A}(a \mid W) = q_{0,A}(a - \delta(W) \mid W)\) for a user-supplied shifting function \(\delta(W)\). Haneuse and Rotnitzky (2013) showed that estimating the effect of this intervention is equivalent with that of an intervention that modifies the value \(A\) would naturally assume according to a regime \(d(A, W)\). Importantly, the regime \(d(A, W)\) may depend on both the covariates \(W\) and the exposure level \(A\) that would be assigned in the absence of the regime; consequently, this has been termed a modified treatment policy (MTP). Both Haneuse and Rotnitzky (2013) and Diaz and van der Laan (2018) considered an MTP of the form \(d(a, w) = a + \delta(w)\) for \(\delta(w) = \gamma \in \mathbb{R}\) if \(a + \gamma \leq u(w)\) and \(d(a, w) = a\) if \(a + \gamma > u(w)\), where \(u(w)\) is the maximum value in the support of \(q_{0,A}(\cdot \mid W = w)\). This intervention generates a counterfactual random variable \(Y_{A+\delta(W)} := f_Y(A + \delta(W), W, U_Y)\) whose distribution we denote \(P_{d}^X\); we seek to estimate \(\psi_{d,\delta} := \mathbb{E}_{P_{d}^X}[Y_{A+\delta(W)}]\), the mean of this counterfactual outcome.

In the context of HVTN 505, this parameter corresponds to the counterfactual 1-year risk of HIV-1 infection had immune response markers of vaccinated participants been increased by \(\gamma\) units relative to the level induced by the current vaccine. This quantity may reflect risk of infection under a next-generation HIV vaccine with improved immunogenicity relative to the vaccine evaluated in HVTN 505. Although the magnitude of shifting could generally be allowed to vary with \(W\), we focus on an intervention that uniformly shifts all participants’ immune
responses by \( \gamma \), that is, \( d(a, w) = a + \gamma \) for all \( a \). Note that for HVTN 505, the parameter of interest is defined only for the vaccine group, making \( A \) a post-vaccination marker measuring an HIV-specific immune response. Importantly, it is not conceivable to define the target parameter for placebo recipients as only HIV-negative participants are enrolled in the trial and \( A \) is only defined if measured prior to HIV infection; consequently, all relevant placebo recipients have value zero for the marker \( A \), and it is not meaningful to apply \( d(a, w) \) to shift the distribution of \( A \).

Analysis of HVTN 505 is complicated by its two-phase design, a technique commonly used for sampling in vaccine efficacy trials. In this sampling scheme, \( X \) is not observed on all participants. Instead, we observe \( O = (W, C, CA, Y) \sim P_0 \), where \( C \) is an indicator that an observation is included in the second-phase sample; \( C_i = 1 \) if \( A \) is measured on the \( i \)th observation and \( C_i = 0 \) otherwise. By convention, \( CA \) denotes that unobserved values of \( A \) are set to zero; this arbitrary labeling has no impact on subsequent developments. For each \( w \) and \( y \), we define \( g_{0,CA}(y, w) := P(C = 1 \mid Y = y, W = w) \), allowing that the probability of inclusion in the second-phase sample can depend on \( W \) and \( Y \). Consequently, the model for \( P_0 \) can be expressed as \( \mathcal{M} = \{P_{P,\delta} : \mathcal{P} \in \mathcal{M}^\delta, g_{0,CA} \} \), that is, \( P_0 \) is implied by the pair \( \{\mathcal{P}^\delta, g_{0,CA} \} \). For example, in HVTN 505 all infected participants with samples available for marker measurement at week 28 had immune responses measured; that is, \( g_{0,CA}(1, w) = 1 \) for all \( w \); however, only a subset of noninfected participants had immune responses measured. We will assume access to an iid sample \( O_1, \ldots, O_n \), denoting its empirical distribution by \( P_n \). We develop efficient nonparametric estimators of \( \psi_{0,\delta} \) based on these data.

### 2.2 Identifying the counterfactual mean under a stochastic intervention

Díaz and van der Laan (2012) established that \( \psi_{0,\delta} \) is identified by

\[
\psi_{0,\delta} = \int_{\mathcal{X}} \int_{\mathcal{A}} \tilde{Q}_{0,Y}(a + \delta(w), w)q_{0,A}(a \mid W = w) \times q_{0,W}(w)d\mu(a)d\nu(w),
\]

where \( \tilde{Q}_{0,Y}(a, w) := E_{P_0^\delta}[Y \mid A = a, W = w] \), the conditional mean of \( Y \) given \( A = a \) and \( W = w \), as implied by \( P_0^\delta \). Let \( Y_{a+i\delta(w)} \) denote the outcome that would have been observed had the observed exposure been, possibly counter-to-fact, set to \( a_i + \delta(w) \); identification of the causal estimand of interest by Equation (1) is established under several assumptions: consistency (\( Y_{i,a_i+\delta(w)} = Y_i \) in the event \( A_i = a_i + \delta(w) \), for \( i = 1, \ldots, n \)); no interference (\( Y_{i,a_i+\delta(w)} \) does not depend on \( a_j + \delta(w) \) for \( i \neq j \) and \( i = 1, \ldots, n \)); no unmeasured confounding (\( A_i \perp \perp Y_{i,a_i+\delta(w)} \mid W = w_i \), for \( i = 1, \ldots, n \)); and positivity (\( a_i \in \mathcal{A} \Rightarrow a_i + \delta(w) \in \mathcal{A} \mid W = w \) for all \( w \in \mathcal{W} \) and \( i = 1, \ldots, n \)). Importantly, even when these untestable assumptions go unsatisfied, the statistical parameter appearing in Equation (1) has a straightforward interpretation: it is the adjusted mean of the outcome \( Y \) under the contrast \( A + \delta(W) \), marginalizing over strata of potential baseline confounders \( W \) (van der Laan and Rose, 2011; Díaz and van der Laan, 2012).

The positivity assumption required to establish Equation (1) is unlike that required for static or dynamic interventions. In particular, it does not require that the post-intervention exposure density place mass across all strata defined by \( W \). Instead, for \( \overline{Q}_{0,Y} \) to be well defined, we require that the density of the exposure mechanism be bounded when the post-intervention exposure mechanism is nonzero, that is, \( 0 < q_{0,A}(A \mid W) \) when \( q_{0,A}(A - \delta(W) \mid W) \neq 0 \), which is satisfied by our choice of \( \delta(W) \).

Díaz and van der Laan (2012) further provided the efficient influence function (EIF) of \( \psi_{0,\delta} \) with respect to a nonparametric model. The EIF evaluated on a typical full-data observation \( x \) can be written

\[
D^{\mathcal{E}}(P_0^\delta)(x) = H(a, w)[y - \overline{Q}_{0,Y}(a, w)] + \overline{Q}_{0,Y}(a + \delta(w), w) - \psi_{0,\delta},
\]

where \( H(a, w) = 1(a < u(w))q_{0,A}(a - \delta(w) \mid w)/q_{0,A}(a \mid w) + 1(a + \delta(w) \geq u(w)) \).

### 2.3 Correcting for two-phase sampling

The subject of two-phase sampling has long been discussed in the statistical literature (Neyman, 1938; Manski and Lerman, 1977; White, 1982). Recent estimation strategies include, among others, methods based on parametric models of the sampling mechanism (Breslow and Cain, 1988), weighted semiparametric estimators (Robins et al., 1994), nonparametric maximum likelihood (Breslow et al., 2003), and re-calibration (Fong and Gilbert, 2015).

Rose and van der Laan (2011) study nonparametric efficiency theory in two-phase sampling designs and provide a representation of the EIF of a target parameter of the full data distribution when the observed data are generated via two-phase sampling. Based on these results,
the EIF in the present problem is
\[
D(G_0, g_{0,C}, D^{F}(p_0^{X}))(o) = \frac{c}{g_{0,C}(y, w)} D^{F}(p_0^{X})(o) \nonumber \\
- \left( \frac{c}{g_{0,C}(y, w)} - 1 \right) G_0(y, w), \tag{3}
\]
where \(D^{F}(p_0^{X})\) is the EIF in Equation (2), and 
\(G_0(y, w) := E_{P_0}[D^{F}(p_0^{X})(O) \mid C = 1, Y = y, W = w].\)

Rose and van der Laan (2011) proposed two estimation strategies. The first—which we call the reweighted estimator—incorporates inverse probability weights based on known or estimated values of the second-phase sampling probability \(g_{0,C}\) to a targeted minimum loss (TML) estimator. The estimator is shown to be asymptotically linear and efficient when \(g_{0,C}\) is known or can be estimated using maximum likelihood. On the other hand, their second estimator can be applied in settings where the sampling design is unknown and must be estimated using non-parametric regression. However, the authors did not provide a formal study of the theoretical properties of this estimator nor numerical evaluations. Owing to its complexity, examples of this approach are limited (eg, Brown, 2014). We aim to fill in these gaps by providing formal theory establishing conditions under which this estimator achieves asymptotic efficiency as well as numerical studies demonstrating its performance in the stochastic intervention context.

### 3 | METHODOLOGY

We utilize two frameworks for estimation: the one-step framework (Pflanzagl and Wefelmeyer, 1985) and TML estimation (van der Laan and Rubin, 2006). Both develop in two stages. In the first stage, we construct initial estimators of key nuisance quantities, whereas in the second stage we perform a bias-correction based on the estimated EIF. The one-step bias correction updates an initial substitution estimator by adding the empirical mean of the estimated EIF, while the TML estimation framework uses a univariate logistic tilting model to build a targeted estimator of \(\tilde{Q}_{0,Y}\) that is subsequently used to construct a plug-in estimator.

#### 3.1 | Estimating nuisance parameters

Our general strategy for estimating nuisance parameters relies on first using the entire observed data set to estimate the second-phase sampling probabilities, \(g_{0,C}\). Subsequently, inverse probability of sampling weights based on these estimates are used to generate estimates of relevant full data quantities using data available only on observations in the second-phase sample. These quantities include the outcome regression \(\tilde{Q}_{0,Y}\), the exposure density \(q_{0,A}\), and the joint distribution of covariates and exposure, which we denote by \(Q_{0,AW}\). Finally, estimates of full data quantities are used to estimate \(G_0\), the conditional mean of the full data EIF given \(Y\) and \(W\) among observations included in the second-phase sample.

Excepting \(Q_{0,AW}\), which we estimate using an inverse probability of sampling weighted empirical distribution, we describe both parametric and flexible, data adaptive estimators. The data adaptive estimators are more parsimonious with our theoretical developments, which permit to nonparametric-efficient estimation; nevertheless, our developments hold equally well for parametric working models. In Theorem 1, we detail assumptions on the stochastic behavior of estimators of these nuisance functions and relate these to the behavior of the resultant estimator of the target parameter.

An estimator of the sampling mechanism \(g_{0,C}\) could be derived from any classification method (eg, logistic regression), in which \(P_{P_0}(C = 1 \mid Y, W)\) is estimated using the full sample; however, nonparametric or semiparametric estimation may be preferable depending on the availability of information about the two-phase sampling design.

To generate an estimate \(Q_{n,AW}\) of the full data joint distribution of \((A, W)\), we use a stabilized inverse probability weighted empirical distribution. For a given \((a, w)\), \(Q_{n,AW}(a, w) := \sum_{i=1}^n C_i / g_{n,C}(Y_i, W_i) 1(A_i \leq a, W_i \leq w) / \sum_{i=1}^n C_i / g_{n,C}(Y_i, W_i)\). To estimate \(Q_{0,Y}\), one may again use any classification or regression model, where \(Y\) is the outcome and functions of \(A\) and \(W\) are included as predictors. In fitting this model, inverse probability of sampling weights \(C_i / g_{n,C}(Y_i, W_i)\) for \(i = 1, \ldots, n\), are included to account for the two-phase sampling design. Any valid regression estimator may be leveraged for this purpose, so long as the implementation of the estimator respects the inclusion of sample-level weights; in practice, we recommend the use of a semiparametric or nonparametric estimator. We denote by \(\tilde{Q}_{n,Y}(a, w)\) the estimate evaluated on a data unit with \(A = a, W = w\).

The simplest strategy for estimating the generalized propensity score \(q_{0,A}\) is to assume a parametric working model and use parametric regression to generate suitable density estimates. Unfortunately, most such approaches do not allow for flexible modeling of \(q_{0,A}\). The relative dearth of available estimators of a conditional density motivated our development of a novel estimator that accounts for two-phase sampling designs. We detail this approach in the Supporting Information and provide an implementation of
our proposal in the haldensify R package (Hejazi et al., 2020). Going forward, we denote by $g_{n,A}(a \mid w)$ the estimated conditional density of $A$ given $W = w$, evaluated at $a \in A$.

The final nuisance parameter that must be estimated is $G_0$, the conditional mean of the random variable $D^F(P_0^X)(O)$ given $(Y, W)$ among those included in the second-phase sample. To estimate this quantity, we generate a pseudo-outcome as follows. First, define the substitution estimator $\psi_{n,\delta} := \int \overline{Q}_{n,Y}(a + \delta(w), w) dQ_{n,AW}(a, w)$, and the auxiliary term $H_n(a, w) := 1(a < u(w))q_{n,A}(a - \delta(w) \mid w)/q_{n,A}(a \mid w) + 1(a + \delta(w) \geq u(w))$. Using these quantities, for all $i$ such that $C_i = 1$, we compute $D_{n,i} := H_n(A_i, W_i)(Y_i - \overline{Q}_{n,Y}(A_i, W_i)) + \overline{Q}_{n,Y}(A_i + \delta(W_i), W_i) - \psi_{n,\delta,i}$. A simple estimation strategy for $G_0$ is to adopt a parametric working model and fit, for example, a linear regression of the pseudo-outcome $D_{n,i}$ on basis functions of $Y$ and $W$. Importantly, since $G_0$ is defined as a conditional expectation with respect to the observed data distribution, we need not include inverse probability of sampling weights in this regression estimate. Although a parametric working model for $G_0$ is permissible, given the complexity of the object, correct specification of this model is likely challenging and we recommend more flexible approaches. We let $G_0(Y_i, W_i)$ denote the value of the chosen regression estimator evaluated on the $i$th observation $i = 1, \ldots, n$.

### 3.2 Efficient estimation

#### 3.2.1 One-step estimator

Based on the nuisance functions described above, efficient estimators may be constructed using either of the one-step or targeted minimum loss estimation frameworks. The one-step estimator adds the empirical mean of the estimated EIF to the initial plug-in estimator, $\psi^+_{n,\delta} := \psi_{n,\delta} + n^{-1} \sum_{i=1}^{n} [C_i / g_{n,C}(Y_i, W_i)D_{n,i} - (C_i / g_{n,C}(Y_i, W_i) - 1)G_n(Y_i, W_i)]$. The resultant augmented one-step estimator $\psi^+_{n,\delta}$ relies on the nuisance functions estimators $(\overline{Q}_{n,Y}, g_{n,A}, G_n, g_{n,C})$. Theorem 1 details sufficient assumptions on these estimators for ensuring that the one-step is asymptotically efficient.

#### 3.2.2 Targeted minimum loss estimator

An asymptotically linear TML estimator of $\psi_{0,\delta}$ may be constructed by using inverse probability of sampling weights to update the initial estimator $\overline{Q}_{n,Y}$ to an estimator $\overline{Q}_{n,Y}$. An updated plug-in estimator is then constructed, $\psi^*_{n,\delta} := \int \overline{Q}_{n,Y}(a + \delta(w), w) dQ_{n,AW}(a, w)$. This updated estimator $\overline{Q}_{n,Y}$ is constructed in a single iteration as follows.

1. Define a working logistic regression model for the conditional mean of $C$ given $(Y, W)$, using the logit of the initial estimate of the censoring mechanism, $\logit(g_{n,C})$, as an offset and with covariate $(G_n/g_{n,C})$. The parameter $\xi \in \mathbb{R}$ corresponding to the covariate may be estimated by maximum likelihood, producing an estimate $\xi_n$. Following estimation of $\xi_n$, this working model yields $g^*_{n,C}$, an updated estimate of the censoring mechanism.

2. Next, define a working logistic regression model for the conditional mean of $Y$ given $(A, W)$, taking the initial estimate of the outcome mechanism $\logit(\overline{Q}_{n,Y})$ as an offset and with covariate $H_n$. The parameter $\epsilon \in \mathbb{R}$ can be estimated via weighted logistic regression (with weights $C_i/g^*_{n,C}(Y_i, W_i)$) to yield an estimate $\epsilon_n$ of $\epsilon$. Using $\epsilon_n$ and this working model, we may update the outcome mechanism to $\overline{Q}_{n,Y}^\epsilon$.

The targeting steps are carried out based on local least favorable parametric submodels, generally requiring only a single iteration for convergence. When the first step of this procedure is omitted, the resultant TML estimator is equivalent to the reweighted estimator of Rose and van der Laan (2011). The additional step allows our estimator to attain asymptotic linearity in a broader set of circumstances. That is, while the reweighted estimator requires that the sampling weights be known or be estimable at a parametric rate, our approach allows for the use of more flexible estimators of sampling weights. Algorithm 1, presented in the Supporting Information, formalizes the proposed procedure.

### 3.2.3 Asymptotic analysis of efficient estimators

We establish the asymptotic efficiency of our estimators in Theorem 1. The theorem depends on a several regularity conditions, which are discussed in the Supporting Information. The theorem is provided in the context of the TML estimator, but, with a similar set of assumptions, the same result holds for the one-step estimator; for brevity, we omit this analogous result. In the sequel, $D^F(\overline{Q}_{0,Y}, q_{0,A})$ and $D^F(P_0^X)$ are used interchangeably as $D^F$ depends on $P_0^X$ only through $\overline{Q}_{0,Y}$ and $q_{0,A}$.
Theorem 1. Asymptotic linearity and efficiency of the TML estimator $\psi_{n,\delta}^*$. Under conditions C1–C6, $n^{-1/2}(\psi_{n,\delta}^* - \psi_{0,\delta}^*) = n^{-1/2} \sum_{i=1}^{n} D(G_0, g_{0,C}, D^2(\hat{Q}_{0,Y}, g_{0,A}))(O_i) + o_p(1)$.

An immediate corollary of theorem 1 is that $\psi_{n,\delta}^*$ is asymptotically efficient, as it is an asymptotically linear estimator with influence function equal to the efficient influence function. Moreover, the central limit theorem implies that the scaled, centered estimator converges in distribution to a mean-zero Gaussian random variable with variance matching that of the EIF (ie, $E_{P_i}\{D^2(G_0, g_{0,C}, D^2(\hat{Q}_{0,Y}, g_{0,A}))(O_i)\}$).

The proof of Theorem 1 is given in the Supporting Information. The conditions of the theorem are standard in semiparametric inference problems, essentially requiring a sub-parametric rate of convergence of each of the nuisance estimators to their true counterparts in terms of an $L^2(P_0)$ norm, a Donsker class condition on the EIF evaluated at the estimated nuisance parameters, and $L^2(P_0)$-consistency of this same object.

In particular, condition C1 requires that the EIF equation be solved and is satisfied by our proposed estimators, while condition C6 requires that the estimated EIF fall in a Donsker class. This latter requirement may be satisfied by using highly adaptive lasso (HAL) regression for all nuisance parameters or avoided entirely by a particular variant of cross-validation (Klaassen, 1987; Zheng and van der Laan, 2011; Chernozhukov et al., 2018). Such an estimator enjoys the same asymptotic properties as our nonsample-splitting estimator while eschewing the Donsker class condition.

Conditions C3–C5 address the behavior of nuisance parameter estimators. Specifically, condition C3 requires that $g_{n,C}$ and $G_0$ converge in $L^2(P_0)$ norm to their true counterparts while condition C4 necessitates negligibility of a second-order remainder term arising from convergence of $q_{0,A}$ and $\hat{Q}_{n,Y}$ to their true counterparts. In the same vein, condition C5 is satisfied by the EIF evaluated at nuisance parameter estimates converging to the EIF evaluated at the true nuisance functions. With respect to these conditions, we note that the HAL regression estimator has been shown to achieve a sufficiently fast rate of convergence so as to satisfy these convergence requirements (van der Laan, 2017; Bibaut and van der Laan, 2019) under the assumption that the true regression function is right-hand continuous with left-hand limits and bounded sectional variation norm. This provided further motivation for our development of a HAL-based conditional density estimator. To increase the applicability of our theorem, our simulation studies and analysis of the HVTN 505 data utilize HAL.

Condition C2 requires that the true sampling mechanism $g_{0,C}$ be bounded away from zero by a (small) positive constant. It is required to ensure that the two-phase sampling procedure does not systematically censor particular strata; the same bound holding for the estimate $g_{n,C}^*$ is required for consistency of the estimator in $L^2(P_0)$ norm. Although many of the conditions of the theorem stipulate that the nuisance parameters converge to their true counterparts in large samples, in finite samples, it may be beneficial to avoid small values of the sampling probability $g_{n,C}$. This can be achieved in an ad hoc way by truncation or more formally via collaborative targeted minimum loss estimation (van der Laan and Gruber, 2010).

3.2.4 Multiple robustness of efficient estimators

The EIF of our estimators enjoys a multiple robustness property, which allows our estimators to achieve consistency even in situations where certain combinations of nuisance parameters are inconsistently estimated.

Lemma 1 (Multiple robustness of the EIF). Let $(G, g_{C}, \hat{Q}_{Y}, q_A)$ denote the limits of the nuisance estimators $(G_0, g_{0,C}, \hat{Q}_{n,Y}, q_{n,A})$ in probability. Suppose either of the following two conditions hold (i) $G = G_0$ and either $\hat{Q}_{Y} = \hat{Q}_{0,Y}$ or $q_A = q_{0,A}$; (ii) $g_{C} = g_{0,C}$ and either $\hat{Q}_{Y} = \hat{Q}_{0,Y}$ or $q_A = q_{0,A}$. Then $\psi_{n,\delta}^* \rightarrow \psi_{0,\delta}^*$.

In the case of the one-step estimator, the initial nuisance function estimates $g_{n,C}$ and $\hat{Q}_{n,Y}$ are used instead. The lemma implies that our efficient estimators will be asymptotically consistent if at least one of $(G_0, g_{0,C})$ and one of $(\hat{Q}_{0,Y}, q_{0,A})$ are consistently estimated.

3.3 Confidence intervals and hypothesis tests

Theorem 1 established the limiting distribution of our efficient estimators. From the limit distribution, inference for either estimator may be attained in the form of Wald-type confidence intervals and corresponding hypothesis tests.

Consider the null and alternative hypotheses $H_0 : \psi_{0,\delta} = 0$ and $H_1 : \psi_{0,\delta} \neq 0$, and denote by $\psi_{n,\delta}$ either the TML estimator $\psi_{n,\delta}^*$ or the one-step estimator $\psi_{n,\delta}^+$. An asymptotic $(1 - \alpha)$ Wald-type confidence intervals is $\psi_{n,\delta} \pm z_{(1-\alpha/2)} \sigma_n / n^{1/2}$, where a $P$-value for a hypothesis test that $\psi_{0,\delta} = 0$ can be computed as $2\{(1 - \Phi(n^{1/2} | \psi_{n,\delta} | / \sigma_n))$, where $\sigma_n$ is the empirical variance of the estimated
EIF, Φ(⋅) is the CDF of the standard normal distribution, and Φ(1−α/2) is the 1−α/2 standard normal quantile.

These procedures are asymptotically justified under the conditions of theorem 1. Importantly, while multiple robustness implies that consistent estimation of ψ_{0,d} is possible under inconsistent estimation of some nuisance parameters, the validity of confidence intervals and hypothesis tests requires consistent estimation of all nuisance parameters.

4 | SIMULATION STUDIES

The proposed estimators were evaluated using two simulation experiments. In the first, we compare our proposed estimators to alternative estimators proposed in the literature. To highlight the benefits offered by our approach over the simple reweighted estimator of Rose and van der Laan (2011), we focus on how estimation of g_{n,C} influences the estimator performance comparing standard logistic regression and all nuisance parameters.

In a second, we evaluate our estimators in a data-generating mechanism inspired by HVTN 505. We compare the relative performance of proposed one-step and TML estimators. These results are detailed in the Supporting Information.

We generated data by drawing covariates W_1 ∼ Normal (3, 1), W_2 ∼ Bernoulli(0.6), and W_3 ∼ Bernoulli(0.3), exposure A ∼ Normal(2(W_2 + W_3), 1), outcome Y | A,W ∼ Bernoulli(expit((W_1 + W_2 + W_3)/3 − A)) and sampling probability, C | Y,W ∼ Bernoulli(expit((W_1 + W_2 + W_3)/3 − Y)). We used this data-generating process to sample n iid observations for n ∈ {100, 400, 900, 1600, 2500} and used the resultant data to estimate the target parameter with each of the estimators considered. This was repeated 1000 times. We considered estimation of ψ_{0,δ} for δ ∈ {−0.5, 0, 0.5}, where the corresponding true values ψ_{0,δ} were {0.501, 0.415, 0.333}.

We compared the reweighted estimators of Rose and van der Laan (2011) to our proposed estimators. For reference, we also present the results of a naive estimator that ignores the two-phase sampling design. In each of these three cases, we consider one-step and TML estimators, giving six estimators in total. Each estimator was constructed by estimating the exposure mechanism g_{n,A} and outcome mechanism Q_{n,Y} via maximum likelihood based on correctly specified parametric models, while g_{n,C} was constructed using either logistic regression or the highly adaptive lasso. Based on theory, we hypothesized that when g_{n,C} is estimated using logistic regression, both the reweighted estimator and our proposed estimator should be asymptotically linear, which would be supported by observing that the bias of the estimators is o_p(n^{-1/2}). On the other hand, when g_{n,C} using the highly adaptive lasso, the reweighted estimators should not achieve asymptotic linearity, while our proposed estimators should. The naive estimators, which make no adjustment for the two-phase sampling design, were expected to perform poorly throughout.

We compared all estimators in terms of their bias (scaled by n^{1/2}), mean squared-error (scaled by n), and coverage of 95% Wald-style confidence intervals. Figure 1 summarizes our findings for the case δ = 0.5; results for δ = 0 and δ = −0.5 are presented in the Supporting Information.

When the sampling mechanism was estimated via a correctly specified parametric model (upper panel), the reweighted and our proposed estimators behave as expected, with low bias and stable MSE. However, the reweighted estimators display MSE increasing with sample size. On the other hand, our proposed estimators achieve nominal coverage. This occurs because the influence function that is basis for the standard error estimates does not include a first-order contribution resulting from estimation of the sampling mechanism resulting in a conservative standard error estimate. Unsurprisingly, the naive estimator performed poorly in all sample sizes, highlighting the importance of accounting for sampling design.

When the sampling mechanism was estimated using HAL (lower panel), the reweighted estimators do not attain linearity, as evidenced by the scaled bias and MSE increasing with sample size. On the other hand, our proposed estimators have small bias and MSE approaching the efficiency bound, thus demonstrating the benefits of the additional effort required to produce our estimators over the simpler reweighted estimators.

Our second simulation study (Supporting Information) showed that the efficient estimators provide reliable performance in a setting similar to HVTN 505. Importantly, this setting incorporates both continuous and binary baseline covariates and a rare outcome (≈5% incidence). We examine the performance of our proposed one-step and TML estimators at a sample size of n = 1400 across δ ∈ {−2, −1.5, −1, 0.5, 0, 0.5, 1, 2} and all nuisance parameters were estimated via HAL. We found that the proposed estimators achieve low bias and MSE, as well as empirical coverage of confidence intervals near the nominal rate. Overall, the TML estimator had slightly better performance than the one-step estimator.

5 | APPLICATION TO THE HVTN 505 TRIAL

HVTN 505 enrolled 2504 HIV-negative participants and randomized participants 1-to-1 to receive an active vaccine or placebo. The 1-year incidence of HIV-1 infection was
FIGURE 1  Comparison of six estimation strategies for $\psi_{0,\delta}$ for $\delta = 0.5$, across 1000 Monte Carlo simulations for each of five sample sizes. The naive estimators do not make use of the estimated sampling mechanism $g_{n,C}$, so their performance is displayed only in the upper panel, in the interest of visual economy. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

about 1.8% per person-year in the vaccine arm and 1.4% per person-year in the placebo arm, during primary follow-up for HIV-1 acquisition (between week 28 and month 24; the same period as was used for assessment of immune correlates). Blood was drawn at the week 26 visit and immune responses measured for all HIV-1 cases diagnosed between week 28 and month 24 and a stratified random sample of uninfected controls (Janes et al., 2017). The two-phase sampling of vaccine-recipient controls sampled five controls per case without replacement within each of eight baseline covariate strata defined by categories of body mass index and race/ethnicity (White, Hispanic, and Black). Janes et al. (2017) and Fong et al. (2018) analyzed these immune responses, and both found CD4+ and CD8+ polyfunctionality scores to be associated with risk of HIV-1 infection status by month 24.

We examined how a range of posited shifts in standardized polyfunctionality scores of the CD4+ and CD8+ immune markers ($A$) would impact the mean counterfactual risk of HIV-1 infection ($Y$) in vaccine recipients. We considered standardized polyfunctionality scores, so that our pre-specified grid of shifts $\delta \in \{-2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0\}$ can be interpreted as shifts on the standard deviation (sd) scale. We present results based on our TML estimator; results for the one-step estimator were similar. To summarize the relationship between the mean counterfactual risk of HIV-1 infection and shifts in polyfunctionality scores, a working marginal structural model (MSM) was constructed, as detailed in Section S3 of the Supporting Information. Our augmented TML estimator $\hat{\psi}_{n,\delta}^*$ requires the construction of initial estimators of all nuisance functions.

The conditional probability of inclusion in the second-phase sample was estimated using HAL, adjusting for age, sex, race/ethnicity, body mass index, and a behavioral risk score for HIV-1 infection. The density $q_{n,A}$ of the CD4+ or CD8+ polyfunctionality scores ($A$), conditional on the same set of covariates ($W$), was estimated using our proposed HAL-based conditional density estimator. The outcome regression $Q_{0,Y}$, which estimated the risk of HIV-1 infection by 24 months ($Y$) given polyfunctionality score and baseline covariates, was estimated using
Inference for the estimates is based on pointwise Wald-type confidence intervals. The slope of a linear working MSM $\hat{\beta}_{\text{TMLE}}$, whose linear form is postulated a priori, summarizes the effect of shifting the polyfunctionality scores on the mean counterfactual risk of HIV-1 infection. A spline model, whose V-shape appears to more closely trace the profile of counterfactual HIV-1 risk changes with $\delta$, was constructed post hoc.

Examination of the point estimates and confidence intervals of $\psi_{0,\delta}$ in Figure 2 reveals that downshifts in the CD4+ polyfunctionality score led to a small increase in estimated HIV-1 infection risk among vaccine recipients (Figure 2, top panel). For example, examining the individual point estimates, a shift of two standard units lower in the CD4+ polyfunctionality score was found to at least double the counterfactual risk of HIV-1 infection. The estimated slope parameter of the working MSM $\hat{\beta}_{\text{TMLE}}$ pointed to an estimated decrease in risk of about $-0.3\%$ per standard unit of CD4+ polyfunctionality change.

The estimated result of shifts in the polyfunctionality score of the CD8+ immunological marker displayed a markedly stronger relationship with the risk of HIV-1 infection.

**FIGURE 2** TML estimates of the counterfactual mean of HIV-1 infection in vaccine recipients under stochastic interventions on CD4+ (top) and CD8+ (bottom) standardized polyfunctionality scores. Inference for the estimates is based on pointwise Wald-type confidence intervals. The slope of a linear working MSM $\hat{\beta}_{\text{TMLE}}$, whose linear form is postulated a priori, summarizes the effect of shifting the polyfunctionality scores on the mean counterfactual risk of HIV-1 infection. A spline model, whose V-shape appears to more closely trace the profile of counterfactual HIV-1 risk changes with $\delta$, was constructed post hoc.
infection (Figure 2, lower panel). Positive shifts of the standardized CD8+ polyfunctionality score beyond those observed in the trial do not appear to have a strong effect on HIV-1 infection risk; confidence intervals for counterfactual risk estimates for all \( \delta \geq 0 \) overlap. On the other hand, shifts that lower the CD8+ polyfunctionality score display a negative linear trend; moreover, confidence intervals for HIV-1 risk estimates at all \( \delta < 0 \) do not overlap with those of the estimate at \( \delta = 0 \); thus, an abundance of caution is warranted in drawing conclusions as to whether shifted CD8+ polyfunctionality score would have improved the HVTN 505 vaccine. Still, it may be informative that the counterfactual HIV-1 infection risk is over four times that observed in the HVTN 505 trial at the largest negative shift considered.

Overall, the results of our analyses support the conclusions of Janes et al. (2017) and Fong et al. (2018), further indicating that modulation of the CD4+ and CD8+ polyfunctionality scores may reduce the risk of HIV-1 infection, with CD8+ polyfunctionality playing a particularly important role. Notably, our analysis differs from the previous efforts in two ways: our estimates (a) are based on a formal causal model, which provides an alternative estimand to summarize relationships between immunogenic response and risk of HIV-1 infection, and (b) leverage machine learning to allow the use of flexible modeling strategies while simultaneously delivering robust inference.

6 | DISCUSSION

A possible criticism of our approach is that, in the context of vaccines, the immune responses we consider may not be directly manipulable. Nevertheless, we believe the estimands that we consider pass the important litmus test question: “If we knew the value of the estimand, could we do something useful with it to advance science?” (Gilbert et al., 2011). Knowledge of which immune responses may lead to the largest decrease in infection or disease incidence would advance vaccine science and stimulate new ideas for the next generation of vaccine research. Another challenge associated with our approach, as with many examples in causal inference, is selecting a scientifically meaningful intervention (ie, modification of the exposure distribution). Although we focused here on additive shifts for simplicity, more scrutiny of this choice is warranted in practice. Scientific context could provide some clues as to potentially meaningful shifts. For example, in the context of influenza vaccination, past studies have shown that repeated vaccinations may have a dulling effect on immune responses to new vaccines (Thompson et al., 2016). When such covariate data are available, they could be used to define an appropriate shift, where the proposed shift is lessened for individuals with many prior vaccinations.

Our analysis of the HVTN 505 trial could be improved in several respects. First, there was participant dropout observed in the trial, which our analysis ignored. A more robust analysis could leverage available covariate information to account for potentially informative missingness. Further, while we investigate the effect of altering post-vaccination immune responses on HIV-1 infection, the issue of interference could limit identifiability of our target causal effects. In trials conducted across geographically diverse sites or within short time frames, it may be reasonable to assume that the potential protection conferred by immune response in a given unit would not alter the infection process in another unit, satisfying the lack of interference requirement for identifiability of the causal parameter of interest. Indeed, there is a growing body of work on relaxing this condition in causal inference (eg, Hudgens and Halloran, 2008).

Beyond this issue, there are several other directions for potentially interesting extensions. First, when a range of shifts is of interest as in our example, we summarized linear trends using working marginal structural models. An alternative formulation could examine the stronger null hypothesis that \( H_0 : E(Y | \delta) = E(Y) \), uniformly in \( \delta \). This is analogous to the hypothesis tests of Kennedy (2019), which deals with shifted binary exposure distributions. There, the authors propose a test of this strong null hypothesis and describe methods for obtaining simultaneous confidence bands using the multiplier bootstrap. Second, it is of interest to extend our estimation strategy to other effects based on stochastic interventions, such as the population intervention (in)direct effects (Diaz and Hejazi, 2020). Extending our estimation strategy to such settings and its application in analyzing other vaccine efficacy trials will be the subject of future research.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this paper.

ORCID

Nima S. Hejazi https://orcid.org/0000-0002-7127-2789
Mark J. van der Laan https://orcid.org/0000-0003-1432-5511
Holly E. Janes https://orcid.org/0000-0002-3237-984X
Peter B. Gilbert https://orcid.org/0000-0002-2662-9427
David C. Benkeser https://orcid.org/0000-0002-1019-8343
REFERENCES

Bibaut, A.F. and van der Laan, M.J. (2019) Fast rates for empirical risk minimization with cadlag losses with bounded sectional variation norm. Preprint, 2019, arXiv:1907.09244.

Breslow, N. and Cain, K. (1988) Logistic regression for two-stage case-control data. Biometrika, 75, 11–20.

Breslow, N., McNeney, B. and Wellner, J.A. (2003) Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. The Annals of Statistics, 31, 1110–1139.

Brown, D.M. (2014). Applications of causal inference to problems of occupational epidemiology. PhD thesis, UC Berkeley.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., et al., (2018) Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21, C1–C68.

Diaz, I. and Hejazi, N.S. (2020) Causal mediation analysis for stochastic interventions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82, 661–683.

Diaz, I. and van der Laan, M.J. (2012) Population intervention causal effects based on stochastic interventions. Biometrics, 68, 541–549.

Diaz. I. and van der Laan, M.J. (2018) Stochastic treatment regimes. In: Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. Amsterdam: Springer Science & Business Media, pp. 167–180.

Fong, Y. and Gilbert, P. (2015) Calibration weighted estimation of semiparametric transformation models for two-phase sampling. Statistics in Medicine, 34, 1695–1707.

Fong, Y., Shen, X., Ashley, V.C., Deal, A., Seaton, K.E., Yu, C., et al., (2018) Modification of the association between t-cell immune responses and human immunodeficiency virus type 1 infection risk by vaccine-induced antibody responses in the HVTN 505 trial. The Journal of Infectious Diseases, 217, 1280–1288.

Gilbert, P.B., Hudgens, M.G. and Wolfson, J. (2011) Commentary on “Principal stratification — a goal or a tool?” by Judea Pearl. The International Journal of Biostatistics, 7, 1341.

Gilbert, P.B., Yu, X. and Rotnitzky, A. (2014) Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. Statistics in Medicine, 33, 901–917.

Hammer, S.M., Sobieszczyn, M.E., Janes, H., Karuna, S.T., Mulligan, M.J., Grove, D., et al., (2013) Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. New England Journal of Medicine, 369, 2083–2092.

Haneuse, S. and Rotnitzky, A. (2013) Estimation of the effect of interventions that modify the received treatment. Statistics in Medicine, 32, 5260–5277.

Hejazi, N.S. and Benkeser, D.C. (2020a). txshift: efficient estimation of the causal effects of stochastic interventions. R package version 0.3.4.

Hejazi, N.S. and Benkeser, D.C. (2020b) txshift: Efficient estimation of the causal effects of stochastic interventions in R. Journal of Open Source Software.

Hejazi, N.S., Benkeser, D.C. and van der Laan, M.J. (2020). haldensity: Conditional density estimation with the highly adaptive lasso. R package version 0.0.6.

Hudgens, M.G. and Halloran, M.E. (2008) Toward causal inference with interference. Journal of the American Statistical Association, 103, 832–842.

Janes, H.E., Cohen, K.W., Frahm, N., De Rosa, S.C., Sanchez, B., Hural, J., et al. (2017) Higher t-cell responses induced by DNA/rAd5 HIV-1 preventive vaccine are associated with lower HIV-1 infection risk in an efficacy trial. The Journal of Infectious Diseases, 215, 1376–1385.

Kennedy, E.H. (2019) Nonparametric causal effects based on incremental propensity score interventions. Journal of the American Statistical Association, 114, 645–656.

Kennedy, E.H., Ma, Z., McHugh, M.D. and Small, D.S. (2017) Nonparametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79, 1229–1245.

Klaassen, C.A. (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. The Annals of Statistics, 1548–1562.

Manski, C.F. and Lerman, S.R. (1977) The estimation of choice probabilities from choice based samples. Econometrica: Journal of the Econometric Society, 1977–1988.

Neyman, J. (1938) Contribution to the theory of sampling human populations. Journal of the American Statistical Association, 33, 101–116.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Pearl, J. and Bareinboim, E. (2014) External validity: from do-calculus to transportability across populations. Statistical Science, 579–595.

Pflanzagl, J. and Wefelmeyer, W. (1985) Contributions to a general asymptotic statistical theory. Statistics & Risk Modeling, 3, 379–388.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89, 846–866.

Rose, S. and van der Laan, M.J. (2011) A targeted maximum likelihood estimator for two-stage designs. The International Journal of Biostatistics, 7, 1–21.

Thompson, M.G., Naleway, A., Fry, A.M., Ball, S., Spencer, S.M., Reynolds, S., et al., (2016) Effects of repeated annual inactivated influenza vaccination among healthcare personnel on serum hemagglutinin inhibition antibody response to A/Perth/16/2009 (H3N2)-like virus during 2010-11. Vaccine, 34, 981–988.

van der Laan, M.J. (2017) A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. The International Journal of Biostatistics, 13, 0097.

van der Laan, M.J. and Gruber, S. (2010) Collaborative double robust targeted maximum likelihood estimation. The International Journal of Biostatistics, 6, 1181.

van der Laan, M.J., Polley, E.C. and Hubbard, A.E. (2007) Super learner. Statistical Applications in Genetics and Molecular Biology, 6, 1309.

van der Laan, M.J. and Rose, S. (2011). Targeted learning: causal inference for observational and experimental data. Berlin: Springer Science & Business Media.

van der Laan, M.J. and Rubin, D. (2006) Targeted maximum likelihood learning. The International Journal of Biostatistics, 2, 1043.

White, J.E. (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology, 115, 119–128.

Zheng, W. and van der Laan, M.J. (2011) Cross-validated targeted minimum-loss-based estimation. In: van der Laan, M.J. and Rose, S. (Eds.) Targeted Learning: Causal Inference for Observational and Experimental Data. Berlin: Springer, pp. 459–474.
SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 3, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library. Although no new data were generated or analyzed in the present work, the results of our reported analyses may be reproduced using the publicly available R code at https://github.com/nhejazi/pub_txshift_biometrics. Throughout our simulation experiments and data analyses, we rely on our txshift and haldensify R packages, available at https://github.com/nhejazi/txshift and https://CRAN.R-project.org/package=haldensify, respectively. The txshift R package implements our proposed efficient estimators of the counterfactual mean outcome under a stochastic intervention, while our haldensify R package provides a non-parametric estimator of the generalized propensity score.

How to cite this article: Hejazi NS, van der Laan MJ, Janes HE, Gilbert PB, Benkeser DC. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. Biometrics. 2021;77:1241–1253. https://doi.org/10.1111/biom.13375