Nonparametric estimation of the causal effect of a stochastic threshold-based intervention

Lars van der Laan\textsuperscript{1,2}, Wenbo Zhang\textsuperscript{3}, Peter B. Gilbert\textsuperscript{2,3}

\textsuperscript{1}Department of Statistics, University of Washington, Seattle, Washington
\textsuperscript{2}Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, Washington
\textsuperscript{3}Department of Biostatistics, University of Washington, Seattle, Washington

Correspondence
Lars van der Laan, Department of Statistics, University of Washington, Seattle, WA 98109.
Email: lvdlaan@uw.edu

Funding information
National Institute of Allergy and Infectious Diseases, Grant/Award Number: 2R37AI054165

Abstract
Identifying a biomarker or treatment-dose threshold that marks a specified level of risk is an important problem, especially in clinical trials. In view of this goal, we consider a covariate-adjusted threshold-based interventional estimand, which happens to equal the binary treatment–specific mean estimand from the causal inference literature obtained by dichotomizing the continuous biomarker or treatment as above or below a threshold. The unadjusted version of this estimand was considered in Donovan \textit{et al.} Expanding upon Stitelman \textit{et al.}, we show that this estimand, under conditions, identifies the expected outcome of a stochastic intervention that sets the treatment dose of all participants above the threshold. We propose a novel nonparametric efficient estimator for the covariate-adjusted threshold-response function for the case of informative outcome missingness, which utilizes machine learning and targeted minimum-loss estimation (TMLE). We prove the estimator is efficient and characterize its asymptotic distribution and robustness properties. Construction of simultaneous 95\% confidence bands for the threshold-specific estimand across a set of thresholds is discussed. In the Supporting Information, we discuss how to adjust our estimator when the biomarker is missing at random, as occurs in clinical trials with biased sampling designs, using inverse probability weighting. Efficiency and bias reduction of the proposed estimator are assessed in simulations. The methods are employed to estimate neutralizing antibody thresholds for virologically confirmed dengue risk in the CYD14 and CYD15 dengue vaccine trials.

Keywords
causal inference, nonparametric efficient estimation, stochastic intervention, targeted minimum-loss estimation, threshold estimation, vaccine trials

1 \quad INTRODUCTION AND PREVIOUS WORK

In clinical trials, it is often of interest to identify a biomarker that is predictive of a clinical outcome of interest. In particular, in vaccine efficacy trials, it is of interest to find so-called correlates of risk, such as neutralizing antibody titer, that are indicative of the risk of acquisition of disease. If additional analyses show that such a biomarker correlate is also a valid surrogate endpoint, then it can be used to predict vaccine efficacy of new vaccines by only analyzing the biomarker, as opposed to observing the clinical endpoints. Such an analysis generally requires only hundreds, rather than thousands, of participants in the vaccine study, and allows for vaccine efficacy to be assessed efficiently in terms of both economic resources and time.
As an intermediate step toward meeting this objective, it is of interest to determine a threshold value of an immune-
response biomarker that predicts a low risk of disease. While there is a large literature of statistical methods for estimating correlates of risk and protection (eg, Chan et al., 2002; Siber et al., 2007; Callegaro and Tibaldi, 2019), prominent methods rely on parametric assumptions, which suffer from misspecification when their strict assumptions are not met. Donovan et al. (2019) proposed a nonparametric minimum-loss estimator for the so-called unadjusted threshold-response function \( E[Y | A \geq v] \) with \( Y \) the outcome and \( A \) the biomarker, and used the nonparametric bootstrap for inferences. The threshold-response function can be viewed as a dose-response-like curve, which maps each threshold to the expected outcome given the biomarker is above that threshold. Unlike the nonparametric dose-response curve, the threshold-response function is \( \sqrt{n} \)-estimable, allowing one to construct efficient estimators and 95% (simultaneous) confidence bands using standard techniques from semiparametric efficiency theory.

Even in randomized trials, the immune-response biomarker, outcome, and outcome missingness mechanisms are generally not randomized itself and thus may be confounded by baseline covariates. In order to have any hope of estimating a causal effect that is due to a threshold-based intervention on the biomarker, covariate adjustment is generally required. Typical techniques such as covariate-stratified estimation, as mentioned in Donovan et al., can lead to confounding bias due to discretization and perform worse as the number of stratification covariates increase. Thus, there is a need to generalize the estimator and estimand proposed by Donovan et al. in a way that allows for flexible and adaptive covariate adjustment. Efficient influence function (EIF)-based causal inference methods (Bickel et al., 1993; van der Laan and Robins, 2003; van der Laan and Rose, 2011) provide general tools for efficient nonparametric estimation of such covariate-adjusted estimands, allowing for the use of machine learning. Specifically, in this article, we employ the targeted learning methodology (van der Laan and Rose, 2011), which provides a general template for constructing efficient substitution estimators.

We extend the previous work of Donovan et al. by allowing for the nonparametric adjustment of arbitrary baseline covariates in the presence of possibly informative outcome missingness. In addition to (1) proposing an efficient targeted minimum-loss estimator (TMLE) that generalizes the estimator of Donovan et al. for the covariate-adjusted threshold-response estimand, \( E_W [E[Y | A \geq v, W]] \), in the presence of possibly informative outcome missingness, further novel additions include the following: (2) establishing rigorous causal identification results for the threshold-response parameter as a stochastic intervention; (3) providing asymptotic efficiency and robustness properties of the proposed TMLE; (4) providing simultaneous confidence bands. Following the treatment by Donovan et al., we discuss in Supporting Information E how to adjust our method when the biomarker is missing at random using inverse-probability weighting. This allows our method to be applied in studies with biased sampling designs.

The covariate-adjusted threshold-response estimand considered in this article has been considered earlier in the literature. Notably, the threshold-response estimand with no outcome missingness is equal to the G-computation-based binary treatment–specific mean estimand (Robins et al., 1994; van der Laan and Robins, 2003; van der Laan and Rose, 2011) where the continuous biomarker or treatment is dichotomized into an indicator of being above or below a threshold. The idea of dichotomizing a continuous biomarker, exposure or treatment variable as an indicator of being above or below a threshold and then applying causal inference methods for binary treatment effects is not a new idea and is used, for instance, in Taubman et al. (2009). Moreover, in an unpublished manuscript, Stitelman et al. (2010) discuss changes in causal interpretation and confounding bias due to discretizing continuous treatments and then applying binary or categorical treatment causal inference methods like the augmented inverse probability weighted (AIPW) or TMLE estimator for the treatment-specific mean (van der Laan and Robins, 2003; Bang and Robins, 2005; van der Laan and Rose, 2011). In sections 4 and 5 of their manuscript, the authors show that the binary treatment–specific mean estimand based on dichotomizing a continuous treatment as above or below a threshold can be viewed as a stochastic intervention. In this article, we provide the following novel additions to the aforementioned authors’ work: (1) We show that in the case of no outcome missingness there is no loss in statistical efficiency by applying binary treatment causal inference methods to estimate the threshold-response estimand; (2) We show that when there is outcome missingness the binary treatment–based estimators are statistically inefficient for the threshold-response estimand, and therefore there is a need to develop a new TMLE that is fully efficient. Moreover, if the outcome missingness is informed by the biomarker value then the binary treatment estimators can be inconsistent due to not fully adjusting for confounding between the continuous treatment and outcome missingness.

The structure of the article is as follows. In Sections 2 and 3, we define the data structure, parameter of interest, and causal assumptions needed for the threshold-response estimand to be identified and interpreted as a stochastic intervention. In Section 4, the EIF of the target parameter is given. In Section 5, we present a novel sequential regression–based TMLE (srTMLE) and establish its
theoretical properties. We also discuss how to construct simultaneous confidence intervals and in the Supporting Information we discuss how to adjust the srTMLE when the biomarker variable is missing at random. In Section 6, we investigate through simulations the efficiency and bias reduction of the proposed srTMLE relative to the Donovan et al. estimator and the inefficient binary treatment-specific mean TMLE (binTMLE). In Section 7, we apply the new method to the CYD14 and CYD15 dengue vaccine trials.

2 NOTATION, DATA STRUCTURE, AND ESTIMAND OF INTEREST

Consider a study where we observe the iid realizations $O_i$ of the random variable $O = (W, A, Δ, ΔY)$ ~ $P_0$, where $P_0$ is the data-generating distribution. Here, $W ∈ ℝ^d$ represents baseline variables, $A ∈ ℝ$ is a continuous biomarker of interest measured during follow-up, $Δ$ is an outcome missingness indicator that takes the value 1 if the outcome is observed, and $Y$ is a binary outcome variable of interest measured at the end of study. We use notation $ΔY$ to denote that $Y$ is observed if and only if $Δ = 1$. For example, in the vaccine trial setting, $A$ may be an immune-response biomarker measured some time during the trial, and $Y ∈ \{0, 1\}$ may be the binary variable that takes the value 1 if the participant acquired the disease study endpoint by the end of the follow-up period, and 0 otherwise. Let $v ∈ ℝ$ be a given threshold in the support of $A$ such that $P_0(A ≥ v \mid W) ≥ δ$ a.e. $W$ for some $δ > 0$, and define the dichotomized biomarker $D_v = 1(A ≥ v)$ and coarsened data-structure $O_v := (W, D_v, Δ, ΔY)$, which will be referenced throughout this article. We assume that $P_0$ is contained in a nonparametric statistical model and let $\| \cdot \|$ denote the $L^2(P_0)$ norm. We will denote $P_W$ as the marginal distribution of $W$, $P_{A\mid W}$ as the conditional distribution of $A \mid W$, $Q(A, W) = E_P[Y = 1 \mid A, W, Δ = 1]$, $Q_0(W) = E_{P_{0}\mid W}[E_P[Y = 1 \mid A, W, Δ = 1] \mid A ≥ v, W]$, $g_{v}(W) = P(A ≥ v \mid W)$. Let $Q_0$, $Q_{0,0}$, $Q_{0,0,0}$, $P_{0,W}$, $P_{0,A\mid W}$ correspond with $P = P_0$. Throughout, we will abbreviate $E_{P_0}, E_{P_{0,A\mid W}}, E_{P_{0,W}}$ as $E_0, E_{0,A\mid W}, E_{0,W}$. Occasionally, we will use the empirical process notation: $P_{n}\hat{f} = E_0f(O)$ and $P_{n}\hat{f} = \frac{1}{n}\sum_{i=1}^{n}f(O_i)$ for a function $o \mapsto f(o)$.

Our estimand of interest is

$$\Psi_{v,0}^{adj} = E_{0,W}[E_0[Y \mid A ≥ v, W)]$$

$$= E_{0,W}(E_{0,A\mid W}[E_0[Y \mid A, W] \mid A ≥ v, W]).$$  \hfill (1)

To ensure the estimand is well defined and identified from the observed data-generating distribution, we will make the following assumptions on $P_0$:

(A1) There exists $δ > 0$ s.t. $P_0(A ≥ v \mid W) > δ$ and $P_0(Δ = 1 \mid A, W) > δ$ a.e. $A, W$.

(A2) $Δ \perp Y \mid A, W$ ($Y$ is missing at random).

Assumption (A1) consists of standard overlap/positivity conditions that ensure the estimands are well defined. Specifically, the assumption ensures that there is a positive probability of observing a biomarker value above the threshold within all strata of $W$ and a positive probability of observing the outcome $Y$ within all strata of $(W, A)$.

Motivated by the latter form of the estimand given in Equation (1), we define the target parameter $ψ_{v}^{adj} : M \mapsto ℝ$:

$$ψ_{v}^{adj}(P) = E_{P_{W}}[E_{P}[Y \mid A, W, Δ = 1] \mid A ≥ v, W]).$$  \hfill (2)

Under assumptions (A1) and (A2), the observed-data estimand $ψ_{v}^{adj}(P_0)$ identifies the estimand given in Equation (1). We will call this estimand the (adjusted) threshold-response function at the threshold $v$, and we will call the map $v \mapsto ψ_{v}^{adj}(P_0)$ the threshold-response function. In the special case where the missingness $Δ$ is not informed by the marker $A$ conditional on $W$ and $A ≥ v$, that is, $Y \perp Δ \mid W, A ≥ v$, Equation (2) reduces to the well-known binary treatment-specific mean estimand: $ψ_{v}^{adj}(P) = E_{P_{W}}[E_{P}[Y \mid A ≥ v, W, Δ = 1]] = E_{P_{W}}[E_{P}[Y \mid D_v = 1, W, Δ = 1]]$. Notably, the simpler form of the estimand is determined by the data-generating distribution of the coarsened data structure $O_v := (W, D_v, Δ, ΔY)$. We will refer to this case as the quasi-informative missingness case, which includes the case of no outcome missingness. We define the unadjusted threshold-response function $ψ_{v}^{unadj}(P) = E_P[Y \mid A ≥ v, Δ = 1]$. In the case where $W$ is independent of $A$, the unadjusted and adjusted threshold response functions are equal.

3 CAUSAL INTERPRETATION AND IDENTIFICATION OF THE THRESHOLD-RESPONSE FUNCTION AS A STOCHASTIC INTERVENTION

Since we have already presented conditions under which the outcome missingness estimand implied by Equation (2) identifies the estimand given in Equation (1), we restrict ourselves to the case of no outcome missingness and consider the estimand given in Equation (1).
Informally and under conditions, the threshold-response estimand can be viewed as the expected outcome under an intervention through the continuous biomarker \( A \) that sets the dichotomized biomarker or binary “treatment” variable \( D_v = 1(A \geq v) \) to 1 for a given individual. Moreover, this intervention is such that an individual with \( A \geq v \) is not intervened upon. From this point of view, the threshold-response estimand is quite similar to standard interventional estimands for binary treatments (Robins et al., 1994; van der Laan and Robins, 2003; van der Laan and Rose, 2011). However, the intervention mechanism associated with the threshold-response estimand acts through the continuous marker \( A \), and, in fact, there are many different interventions on individuals with \( A < v \) preintervention that set \( D_v \equiv 1 \). Since the distribution of the outcome \( Y \) generally depends on the continuous values of \( A \) and not only the binary indicator \( D_v \), different interventions that set \( D_v \equiv 1 \) can lead to different expected counterfactual outcomes (Stitelman et al., 2010).

To rigorously formulate the notion of an intervention, we consider the following causal model. Let \( U_W, U_A, U_Y \) be exogenous random variables, and let \( f_W, f_A, f_Y \) be deterministic functions. We define the underlying causal data structure to be \( O_{causal} = (W, A, Y, U_W, U_A, U_Y) \sim P_{0,2} \). We assume that \( O = (W, A, Y) \) is causally generated by the following nonparametric structural equations model (NPSEM) (Pearl, 2009),

\[
W = f_W(U_W), \quad A = f_A(W, U_A), \quad Y = f_Y(A, W, U_Y).
\]

Following the formulation given in Pearl (2009) and the notation of Diaz et al. (2021), we define an intervention on \( A \) as a rule that maps a realization \((w, a)\) of \((W, A)\) to a possibly randomized biomarker value \( d_0(w, a) \in \text{support}(A) \) that may depend on \( P_0 \). For a given observation \( O = (W, A, Y) \), the counterfactual outcome associated with the intervention \( d_0 \) is \( Y_{d_0} := f_Y(d_0(W, A), W, U_Y) \). If \( A \in \{0, 1\} \) were binary then \( d_{int}(a) := 1(a \geq 1) \equiv 1(a = 1) \) is a valid intervention that corresponds with the standard binary treatment intervention and \( Y_1 := Y_{d_{int}} \) would be the familiar counterfactual outcome associated with the treatment assignment \( A \equiv 1 \).

The threshold-response estimand given in Equation (1) can be viewed as the expected outcome under a randomized intervention \( \tilde{d}_{0,v} \) on \( A \) that satisfies \( 1(\tilde{d}_{0,v}(A, W) \geq v) = 1 \) and is partially stochastic conditional on \((A, W)\). Specifically, \( \tilde{d}_{0,v} \) is given by

\[
\tilde{d}_{0,v}(w, a) := 1(a < v) \cdot F_{0,v}^{-1}(z \mid w) + 1(a \geq v) \cdot a,
\]

where \( F_{0,v} \) is the conditional CDF \( s \sim P_0(A \leq s \mid W = w, A \geq v) \) and \( Z \sim P_Z \) is a randomizer that is uniformly distributed on \([0,1]\) and independent of \( O \). Noting that \( F_{0,v}^{-1}(z \mid w) \) is exactly an independent draw from the conditional CDF \( F_{0,v} \), we are guaranteed that the assigned interventional marker value is always above the threshold \( v \).

Remark 1. In the special case where \( W \) is discretely valued, the above interventional value \( \tilde{d}_{0,v}(w, a) \) can be computed in a more intuitive way as follows. Consider an infinitely large target population consisting of realizations of \((W, A, Y) \sim P_0 \). For a given individual \( O = (W, A, Y) \) with \((W = w, A = a) \) and \( a < v \), uniformly-at-random match this individual with a member \( O_{match,v}(w) \) of the population with the same baseline covariates \( W_{match,v} = w \) but with biomarker value \( A_{match,v}(w) \geq v \). After matching, assign the interventional biomarker value of the individual \( O \) to the observed biomarker value of the matching individual \( O_{match,v}(w) \). For individuals in the sample with \( A \geq v \), no intervention is performed. In this case, the random interventional value can be written as \( \tilde{d}_{0,v}(A, W) = A \cdot 1(A \geq v) + A_{match,v}(W) \cdot 1(A < v) \).

Back to the general case, the expected outcome under the intervention \( \tilde{d}_{0,v} \) is given by \( \Psi_{v, causal}(P_{0,2}) := E_{P_0 \times P_2}[Y_{\tilde{d}_{0,v}(W, A), W, U_Y}] = E_{P_0 \times P_2}[f_Y(\tilde{d}_{0,v}(W, A), W, U_Y)] \), and it is identified from \( P_0 \) by the estimand given in Equation (1) under the first part of Assumption (A1) and the following additional assumption.

\[
(A3) \quad U_A \perp U_Y \mid W (\text{or} Y \mid \{A = a, W\} =_{d_i} Y_a \mid \{A = a, W\} & Y_a \perp A \mid W \text{ for all } a \geq 0).
\]

Viewing \( P(A \geq v \mid W) \) as a threshold-specific propensity score, the first part of Assumption (A1) is analogous to the overlap assumption needed for the identification of binary treatment causal effects (Rubin, 2000; Pearl, 2009). Notably, the overlap condition is fairly mild relative to those for other stochastic interventions on continuous variables (see, eg, Diaz et al., 2021). It only requires that there is some mass assigned to the interval \([v, \infty)\) for all strata of \( W \) and no conditions are imposed on how this mass is distributed. Assumption (A3) requires that there are no unmeasured confounders between \( A \) and \( Y \) and both versions are similar to the assumptions sufficient for identification of other stochastic interventions (Diaz et al., 2021).

## 4 \quad \text{EIF OF THE PARAMETER OF INTEREST AND INEFFICIENCY OF BINARY TREATMENT–BASED ESTIMATORS}

A key object necessary for constructing efficient estimators for an estimand is the EIF, which is uniquely defined given a statistical model and target parameter. The EIF is essential for characterizing the asymptotic distribution of an efficient estimator. Specifically, given the
data-generating distribution $P_0$ contained in the nonparametric statistical model $\mathcal{M}$, the parameter $\Psi_v^{adj}$, and its EIF $D_{P,v}$ indexed by $P \in \mathcal{M}$ and a threshold $v \in \mathbb{R}$, the optimal $\sqrt{n}$-scaled and centered asymptotic distribution among all asymptotically linear and regular (w.r.t. $\mathcal{M}$) estimators of $\Psi_v^{adj}(P_0)$ is mean-zero normally distributed with variance $E_vD_{P,v}(O)^2$. For a more detailed account of the theory of EIFs and semiparametric efficiency theory, we refer to Bickel et al. (1993). The derivation of the EIFs can be found in Supporting Information A.

Lemma 1. The EIF of the parameter $\Psi_v^{adj}$ is
\[
D_{P,v}(W, A, \Delta, \Delta Y) = \frac{1(A \geq v)}{P(A \geq v | W)} \frac{\Delta}{P(\Delta = 1 | A, W)} \times (Y - E_P[Y | A, W, \Delta = 1])
+ \left( E_P[Y | A, W, \Delta = 1] - E_{P,v}[E_P[Y | A, W, \Delta = 1] | A \geq v, W] \right)
\times \frac{1(A \geq v)}{P(A \geq v | W)}
+ E_{P,v}[E_P[Y | A, W, \Delta = 1] | A \geq v, W] - \Psi_v^{adj}(P).
\]

When there is no outcome missingness, that is, $\Delta \equiv 1$, the EIF reduces to the EIF for the binary treatment–specific mean $E_{P,v}E_P[Y | D_v = 1, W]$ where $D_v = 1(A \geq v)$ plays the role of the binary treatment (van der Laan and Robins, 2003; Bang and Robins, 2005; van der Laan and Rose, 2011). For the case of no outcome missingness, this establishes that an estimator for the threshold-response estimand that is nonparametric-efficient w.r.t. the coarsened data-structure $(W, D_v, Y)$ is also efficient with w.r.t. the observed data-structure $(W, A, Y)$. On the other hand, when there is outcome missingness that is only informed by $W$ and not $A$ (so that the estimand reduces to $E_{P,w}E_P[Y | A \geq v, W, \Delta = 1]$), the first two terms of the EIF of Lemma 1 does not reduce to the relevant EIF component for the coarsened data structure $(W, D_v, \Delta, \Delta Y)$, given by $\frac{\Delta(A \geq v)}{P(A \geq v | W)P(\Delta = 1 | W)} \{Y - E_P[Y | A \geq v, W, \Delta = 1]\}$, which implies that estimators based on the coarsened data-structure $O_v = (W, D_v, \Delta, \Delta Y)$ are statistically inefficient with respect to the nonparametric statistical model $\mathcal{M}$. The loss in efficiency of the binary treatment methods is largely driven by how much more $E_P[Y | A, W, \Delta = 1]$ is predictive of $Y$ than $E_P[Y | A \geq v, W, \Delta = 1]$ among individuals with $A \geq v$ (Moore and van der Laan, 2009).

5 | METHODOLOGY

For efficient estimation of the threshold-response function, we employ targeted minimum-loss based estimation (TMLE) (van der Laan and Rose, 2011). TMLE is a two-step approach for constructing nonparametric and often double-robust efficient substitution estimators for pathwise-differentiable parameters of interest. TMLE is closely related to the one-step and estimating equation methodology (Bickel et al., 1993; van der Laan and Robins, 2003; Bang and Robins, 2005) in that it utilizes the EIF in its estimation procedure, but it does so in a way that ensures the resulting estimator is a substitution estimator, which can lead to additional robustness and improved performance in certain finite-sample settings (Porter et al., 2011). In short, given an initial estimator $P_{n,0} \equiv (P_{n,W}, Q_{n,V}, g_{n,V}, G_n)$ of $P_0 \equiv (P_0,W, Q_0, V, G_{0,V}, G_0)$, the TMLE procedure performs minimum-loss estimation with loss $(P, O) \leftrightarrow L(P, O)$ over a data-adaptive parametric submodel $P_{n,e}$ through $P_{n,0}$ that points in the direction of maximal change of the parameter $\Psi_v^{adj}$. We call the resulting risk minimizer $P^*_n$, the TMLE of $\Psi_v^{adj}$. The submodel and choice of loss function have the key property that the “likelihood” scores of the risk function along the submodel equals the empirical mean of the EIF of $\Psi_v^{adj}$. That is, $\frac{d}{dz}P_n L(P_{n,e}) = P_n D_{n,e,v}$. As a consequence, the TMLE, which is the risk minimizer along this submodel, necessarily solves the efficient score equation: $P_n D_{n,e,w} = 0$. This implies $\Psi_v^{adj}(P_{n,e}) = \Psi_v^{adj}(P_n) + P_n D_{n,e,v}$, and therefore the substitution estimator using $P_{n,e}$ equals the one-step efficient estimator (Bickel et al., 1993) and inherits its asymptotic optimality properties. We refer to Gruber and van der Laan (2009) and van der Laan and Rose (2011) for a more thorough introduction and overview of TMLE.

5.1 | Targeted minimum loss estimator

We present a sequential-regression estimator (srTMLE) for the threshold-response estimand that applies in the presence of informative outcome missingness. We also provide an inefficient TMLE (binTMLE), which is a consistent estimator in the case of quasi-informative outcome missingness: $A \perp \Delta | W$. The binTMLE is equal to the binary treatment–specific mean TMLE (van der Laan and Rose, 2011) where the binary treatment is $D_v = 1(A \geq v)$, and is therefore also asymptotically equivalent to the analogous one-step efficient AIPW estimator (see, for instance, van der Laan and Robins, 2003; Bang and Robins, 2005). The binTMLE is statistically inefficient when the missingness is informed by $W$ and inconsistent when informed by $W$ and $A$. In Supporting Information F, we discuss how our proposed estimators can be adjusted when there is missingness (e.g., due to biased sampling) in the biomarker $A$. These methods can also be applied with minor
5.1.1 A novel efficient sequential-regression–based TMLE

The sequential-regression TMLE (srTMLE) requires initial estimation of the nuisance parameters \( Q_0(a, w) = E_0[Y | A = a, W = w, \Delta = 1], \) \( Q_0(v) = E_0[A | Q_0(A, W) | A \geq v, W = w], \) \( G_0(a, w) = P_0(\Delta = 1 | A = a, W = w), \) and \( g_0(v) = P_0(V = v | A = a, W = w). \) Denote their respective estimators \( \hat{Q}_n, \hat{G}_n, \) and \( \hat{g}_n, \) and let \( P_{W, n} \) be the empirical estimator of the marginal distribution of \( W. \) Our initial estimator of (the relevant parts of) \( P_0 \) is given by \( P_{n, v, \nu} := (P_{W, n}, g_{n, \nu}, Q_n, G_n). \) The srTMLE is defined as follows:

1. Define the indicator fluctuation submodel \( Q_{n, \Delta}(A, W) = \expit(\log(Q_n)(A, W) + \epsilon(A \geq v)). \)
2. The MLE along this submodel is given by

\[
\hat{\epsilon}_n = \arg\max_{\epsilon \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{g_{n, \nu}(W_i)g_n(A_i, W_i)} \times \{ Y_i \cdot \log \hat{Q}_n(A_i, W_i) + (1 - Y_i) \times \log(1 - \hat{Q}_n(A_i, W_i)) \}.
\]

3. Define the updated estimate of \( Q_0 \) as \( \hat{Q}_n = Q_{n, \hat{\epsilon}_n}. \)
4. Obtain an initial estimator \( Q_{n, \nu}(W) = E_0[E_0[Y | A, W, \Delta = 1] | A \geq v, W] \) using sequential regression (eg. estimate \( E[Q_0(A, W) | A \geq v, W]). \)
5. Define the intercept fluctuation submodel, \( Q_{n, \nu, \nu} = \expit(\log(Q_{n, \nu}) + \epsilon). \)
6. The MLE along this submodel is given by \( \hat{Q}_{n, \nu} = Q_{n, \nu, \hat{\epsilon}_n} \) where \( \hat{\epsilon}_n = \arg\max_{\epsilon \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \frac{1(A_i \geq v)}{g_{n, \nu}(W_i)} \{ \hat{Q}_n(A_i, W_i) \log \hat{Q}_{n, \nu}(W_i) \\ + (1 - Q_n(A_i, W_i)) \log(1 - \hat{Q}_{n, \nu}(W_i)) \}. \)
7. Let \( P_{n, \nu, \nu} := (P_{W, n}, g_{n, \nu}, Q_n, Q_{n, \nu}, Q_n). \) The TMLE of \( \psi_{\nu, \nu}(P_0) \) is then given by the substitution estimator \( \psi_{\nu, \nu}(P_{n, \nu, \nu}) = E_{P_{W, n}} \hat{Q}_{n, \nu}(W) = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_{n, \nu}(W_i). \)

Step (2) (resp. (6)) requires performing the inverse probability weighted (IPW)-weighted logistic regression of \( Y \) (resp. \( Q_n(A, W) \)) on \( 1(A \geq v) \) (resp. (1)) with offset being an initial estimator of \( Q_n(A, W) \) (resp. \( Q_{n, \nu}(W) \)). In step (5), the initial estimator of \( Q_0, \) can be obtained by treating \( Q_n(A, W) \) as a pseudo-outcome and then performing the nonparametric regression of \( Q_n(A, W) \) on \( W \) using only the observations \( O_1 \) with \( A_i \geq v. \)

The key property of the srTMLE is that the targeted estimators \( \hat{Q}_n \) and \( \hat{Q}_{n, \nu} \) solve the following score equations:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{1(A_i \geq v)}{g_{n, \nu}(W_i)g_n(A_i, W_i)} \{ Y_i - Q_n(A_i, W_i) \} = 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \frac{1(A_i \geq v)}{g_{n, \nu}(W_i)} \{ Q_n(A_i, W_i) - Q_{n, \nu}(W_i) \} = 0,
\]

which implies that the efficient score equation is solved: \( P_0 D_{P_0, \nu} = 0. \)

5.1.2 Inefficient TMLE for quasi-informative missingness

The inefficient binary treatment–based TMLE (binTMLE) is consistent, but inefficient, if \( A \perp \Delta | W, A \geq v. \) We use the same notation as in the previous section except let \( G_{n, v}(w) \) be an estimator of \( P_0(\Delta = 1 | A \geq v, W = w) \) and \( Q_{n, v}(w) \) be an estimator of \( E_0[Y | A \geq v, W, \Delta = 1] \) (which equals \( Q_{n, \nu} \) under the assumption that \( A \perp \Delta | W, A \geq v). \)

1. Define the intercept fluctuation submodel \( Q_{n, \nu, \nu} = \expit(\log(Q_{n, \nu}) + \epsilon). \)
2. Define the MLE along this submodel \( Q_{n, \nu} = Q_{n, \nu, \hat{\epsilon}_n} \) by

\[
\hat{\epsilon}_n = \arg\max_{\epsilon \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \frac{1(A_i \geq v)}{g_{n, \nu}(W_i)g_n(A_i, W_i)} \times \{ Y_i \cdot \log \hat{Q}_{n, \nu}(W_i) + (1 - Y_i) \cdot \log(1 - \hat{Q}_{n, \nu}(W_i)) \}.
\]
3. Let \( P_{n, \nu, \nu} := (P_{W, n}, g_{n, \nu}, Q_n, Q_{n, \nu}, Q_n). \) The binTMLE for \( \psi_{\nu, \nu}(P_0) \) is given by the substitution estimator \( \psi_{\nu, \nu}(P_{n, \nu, \nu}) = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_{n, \nu}(W_i). \)

5.2 Asymptotic inference with the srTMLE

In this section, we present a general theorem that characterizes the asymptotic behavior of the srTMLE defined in Section 5.1.1. An analogous result for the binTMLE defined in Section 5.1.2 follows from van der Laan and Rose (2011). Let \( P_{n, \nu, \nu} := (P_{W, n}, g_{n, \nu}, Q_n, Q_{n, \nu}, Q_n) \) denote the targeted nuisance estimates for the srTMLE of \( \psi_{\nu, \nu}(P_0). \) Let \( K \subset \mathbb{R} \) be a compact set. In order for the srTMLE to be asymptotically linear and efficient, we require the following regularity conditions on the initial estimators and nuisance functions.

- \( B_1: \delta < P_0(Y = 1 | A \geq v) < 1 - \delta \) for some \( \delta > 0. \)
- \( B_2: g_{0, \nu}, G_0 > \delta \) and \( g_{n, \nu}, G_n > \delta \) with probability tending to 1 for some \( \delta > 0. \)
The set of realizations of \( w \mapsto g_{n,v}(w), (a,w) \mapsto G_n(a,w), (a,w) \mapsto Q_n(a,w), \) and \( w \mapsto Q_{n,v}(w) \) are \( P_0 \)-Donsker.

\[
\begin{align*}
\|Q_n - Q_0\| &= o_p(n^{-1/4}), \\
\|Q_{n,v} - Q_{0,v}\| &= o_p(n^{-1/4}), \\
\|g_{n,v} - g_{0,v}\| &= o_p(n^{-1/4}), \\
\|G_n - G_0\| &= o_p(n^{-1/4}).
\end{align*}
\]

The union over all \( v \in K \) of the set of realizations of the functions given in B3 is \( P_0 \)-Donsker.

Condition B1 requires that the random outcome \( Y \) is not degenerate on the event \( \{A \geq v\} \), which ensures that the EIF is nonvanishing. This condition is required for the srTMLE to have an asymptotic distribution after centering and scaling by \( \sqrt{n} \) but is not required for \( \sqrt{n} \)-consistency of the srTMLE. Condition B2 is a standard positivity assumption needed for the estimand of interest and srTMLE procedure to be well defined. Condition B3 requires that the estimators of the nuisance parameters are well behaved with their realizations falling in a not-too-complex function space. More aggressive algorithms like random forests and gradient boosting are prone to overfitting when not properly tuned, which can lead to a violation of this condition. However, by cross-fitting the nuisance estimators and employing CV-TMLE (van der Laan and Rose, 2011; Chernozhukov et al., 2018), this condition can be removed entirely, allowing one to safely employ such estimators (see references for more discussion). Condition B4 requires the nuisance estimators converge fast enough to the nuisance parameters. Both Conditions B3 and B4 are satisfied under smoothness conditions by a number of estimators including risk minimizers over reproducing kernel Hilbert spaces (RKHS), neural networks with VC dimension that does not grow too fast with sample size (Farrell et al., 2018), generalized additive models, and the highly adaptive lasso estimator (Benkeser and van der Laan, 2016). Condition B5 ensures that the nuisance estimator realizations and nuisance parameters for all thresholds \( v \in K \) fall in a single controlled function class. This condition is trivially satisfied for the nuisance estimators whenever Condition B3 is satisfied and \( K \) is finite. Otherwise, this can usually be enforced by pooling the estimation across the thresholds and then employing a machine-learning algorithm with \( P_0 \)-Donsker realizations (eg, use pooled logistic/linear regression). Alternatively, this can be enforced by ensuring that all nuisance estimators for each \( v \in K \) fall in a single Donsker function class (eg, functions of bounded variation or an RKHS with uniformly bounded Hilbert space norm). This can be guaranteed by estimating the nuisance functions with the highly adaptive lasso estimator (Benkeser and van der Laan, 2016) as long as the nuisance functions have a uniformly bounded variation norm.

Theorem 1. Suppose Conditions B1, B2, B3, and B4 hold. Then, the srTMLE estimator \( \Psi^{ad}_{v}(P_{n,v}) \) satisfies

\[
\sqrt{n}(\Psi^{ad}_{v}(P_{n,v}) - \Psi^{ad}_{v}(P_0)) = n^{-1/2} \sum_{i=1}^{n} D_{P_{n,v}}(W_i, A_i, \Delta_i, Y_i) + o_p(1).
\]

If only B1 is violated then one has \( \sqrt{n}(\Psi^{ad}_{v}(P_{n,v}) - \Psi^{ad}_{v}(P_0)) = o_p(n^{-1/2}) \). If in addition Assumption B4 holds uniformly for all \( v \in K \) \( \sqrt{n}(\Psi^{ad}_{v}(P_{n,v}) - \Psi^{ad}_{v}(P_0)) \) converges to a tight mean-zero Gaussian process in \( l^\infty(K) \) with covariance function \( \rho(v_1, v_2) = P_0D_{P_{0,v_1}}D_{P_{0,v_2}} \).

It follows immediately from the preceding theorem that the srTMLE is an efficient estimator for \( \Psi^{ad}_{v}(P_0) \), since it is asymptotically linear with influence function being the EIF. The srTMLE’s scaled and centered asymptotic distribution is given by a mean-zero normally distributed random variable with variance being the variance of its influence function. An estimate \( \sigma_{n,v} \) of the standard error \( \sigma_v : = \sqrt{P_0D_{P_{0,v}}^2} \) of the srTMLE is given by \( \sigma_{n,v}^2 = \frac{1}{n} \sum_{i=1}^{n} D_{P_{n,v}}(W_i, A_i, Y_i)^2 \). Under Condition B3, one has that \( D_{P_{n,v}} \) falls in a class of \( P_0 \)-Donsker functions, which when paired with Condition B4 and the fact that \( D_{P_{n,v}} \) is bounded under Condition B2, implies that \( |\sigma_v - \sigma_{n,v}| = o_p(1) \). Using this estimate of the standard error, Wald-type confidence intervals can be constructed for inference. Specifically, \( \Psi^{ad}_{v}(P_{n,v}) \pm \Phi^{-1}(1 - \alpha)\frac{\sigma_{n,v}}{\sqrt{n}} \) forms an asymptotic \( 1 - \alpha \) confidence interval for \( \Psi^{ad}_{v}(P_0) \) where \( \Phi \) is the CDF of an \( N(0,1) \) random variable.

5.3 Robustness properties of the srTMLE estimator

The srTMLE is double robust with respect to the nuisance estimators. The double-robustness property is due to the structure of the efficient influence function and is analogous to the double-robustness property for the well-known AIPW estimator of the average treatment effect (Bang and Robins, 2005).

Theorem 2. Under Conditions B2, B3, the srTMLE given in Section 5.1.1 is a consistent estimator for \( \Psi^{ad}_{v}(P_0) \) if either of the following conditions hold. If Condition B4 holds as well then the srTMLE is consistent uniformly in \( v \in K \).

- \( P_0(A \geq v \mid W) \) and \( P_0(\Delta = 1 \mid A, W) \) are estimated consistently.
\( E_{0|W}[E_0[Y \mid A, W, \Delta = 1] \mid A \geq v, W] \) and \( E_0[Y \mid A, W, \Delta = 1] \) are estimated consistently.

5.4 Simultaneous confidence bands for the threshold-response function

Let \( V = \{v_1, \ldots, v_k\} \subset K \subset \mathbb{R} \) be a finite set of thresholds for some \( v \in \mathbb{N} \), contained in a bounded set \( K \). In practice, \( V \) represents a discrete grid of threshold values that are of interest. Note that since \( V \) is discrete, we can apply Theorem 1 with \( K = V \). Thus, we have that the collection of srTMLEs \( \{\Psi^{adj}_v(P_{n,v}^{*}) : v \in V\} \) satisfies \( \{\sqrt{n}(\Psi^{adj}_v(P_{n,v}^{*}) - \Psi^{adj}_v(P_0)) : v \in V\} \rightarrow \{Z_v : v \in V\} \), where \((Z_v : v \in V)\) is a mean-zero multivariate normally distributed random variable with covariance matrix \( \Sigma_{v_1,v_2} = P_0D_{P_0,v_1}D_{P_0,v_2} \). The covariance matrix can be estimated consistently with the empirical covariance matrix \( \hat{\Sigma}_{n,v_1,v_2} = P_nD_{P_n,v_1,v_1}D_{P_n,v_2,v_2} \), which further gives a consistent estimate of the distribution of \((Z_v : v \in V)\). Simultaneous confidence intervals for multivariate normally distributed random variables are well understood and we refer to Cai and van der Laan (2020) for an in-depth treatment in the context of TMLE. By the second statement of Theorem 1, the inference remains asymptotically valid even as we let the number of thresholds contained in \( V \) grow and approach the set \( K \). Thus, by taking a sufficiently fine grid \( V \) of thresholds, we can interpolate the estimates and confidence bands of the thresholds in \( V \) to thresholds in \( K \cap V \) at negligible cost in bias and coverage. We note that since \( v \rightarrow \Psi^{adj}_v(P_{n,v}^{*}) \) is a locally efficient, asymptotically linear and regular estimator, these confidence intervals are nonadaptive. In fact, by the local asymptotic minimax theorem (van der Vaart, 1998), any adaptive estimator of \( v \rightarrow \Psi^{adj}_v(P_0) \) that achieves better performance than the efficient estimator at some distribution that is a \( n^{-1/2} \)-fluctuation from \( P_0 \) must necessarily be suboptimal at some distributions, and therefore is not locally efficient. This is in contrast with constructing confidence intervals for general regression functions for which no such efficiency theory exists (Genovese and Wasserman, 2008).

6 SIMULATIONS

6.1 Asymptotic efficiency gains of srTMLE relative to binTMLE

In this section, we explore the efficiency claims made in Section 4. Specifically, we investigate the efficiency gains from using the srTMLE as opposed to the binTMLE. For \( \text{const} \in \{0,1\} \) and \( \text{offset} \in \{0,-3\} \), we consider the following simulation setting: \( A \sim \text{truncnorm}(a = 0, b = 2, \text{mean} = (0.8 + W_1 + (W_2 + W_3)/2)/2, \text{sd} = 0.5) \), \( Y \sim \text{Bern}(\expit(0.75 \cdot W_1 - 0.2 + 0.5 \cdot (W_2 + W_3) - A + \text{const} \cdot 2 \cdot \sin(6 \cdot A))) \), \( \Delta \sim \text{Bern}(\expit(-1 + W_1 + W_2 + W_3)) \) (on average 40% missing) where \( W_1, W_2, W_3 \) are baseline variables (see Supporting Information D). For \( \text{const} = 1 \), the addition of the term \( 2 \cdot \sin(6 \cdot A) \) in the distribution of \( Y \) ensures there is a nonlinear association between \( A \) and \( Y \) that will not be captured well by \( v \rightarrow E_0[Y \mid A \geq v, W] \); \( \text{offset} = 0 \) corresponds with a nonrare event setting with \( P_0(Y = 1) \approx 0.45 \) and \( \text{offset} = -3 \) corresponds with a rare-event setting with \( P_0(Y = 1) \approx 0.08 \). For a grid of 10 thresholds in \( [0,2] \), we estimated (using a large simulated sample) the relative loss in efficiency, defined as \( \frac{\text{sd}(D_{P_0,v}^{\text{coarse}})}{\text{sd}(D_{P_0,v}^{\text{adj}})} - 1 \) where \( D_{P_0,v}^{\text{coarse}} \) is the inefficient influence function based on the coarsened data structure \( (W, D_v, \Delta, \Delta Y) \). \( \text{sd}(D_{P_0,v}^{\text{coarse}}) \) is the asymptotic standard error of the binTMLE, and \( \text{sd}(D_{P_0,v}^{\text{adj}}) \) is the asymptotic standard error of the srTMLE. The results for all settings are displayed in Figure 1. Both the rare and nonrare event settings show that there is a noticeable efficiency loss for the case where \( E_0[Y \mid A, W] \) is nonlinear. This makes sense because \( E_0[Y \mid A, W] \) is more predictive of \( Y \) \( (r = 0.60, r_{\text{rare}} = 0.35) \) than \( E_0[Y \mid A \geq v, W] \) \( (r = 0.08, r_{\text{rare}} = 0.07 \) for \( v = 0 \) (see the discussion in Section 4). The loss in efficiency is much more substantial in the non–rare-event setting, which can be explained by the fact that in the rare-event setting both \( E_0[Y \mid A, W] \) and \( E_0[Y \mid A \geq v, W] \) are usually small and are therefore poor predictors of \( Y \). Next, we see that in both linear settings, there is little to no loss in efficiency. This can be explained by (1) the decrease in the predictive power of \( E_0[Y \mid A, W] \) that is directly due to \( A \) by omitting the nonlinear term \( (r = 0.24, r_{\text{rare}} = 0.1) \) and (2) the monotone relationship between \( E_0[Y \mid A, W] \) and \( A \) that allows for \( E_0[Y \mid A \geq v, W] \) to be more predictive \( (r = 0.15, r_{\text{rare}} = 0.07 \) for \( v = 0 \) than it was in the nonlinear case.

6.2 Asymptotic bias of inefficient (coarsened data structure) TMLE

In this section, we investigate how the following estimators perform in a setting with outcome missingness that is informed by both \( W \) and \( A \): (1) the proposed efficient srTMLE, (2) the inefficient binTMLE, (3) the estimator of Donovan et al. All nuisance functions were estimated using generalized additive models using the function “gam” provided in the R package “mgcv” with default settings. For the distribution of \( A \) and \( Y \), we utilize the same distribution as the previous section with \( \text{offset} = -3 \) and \( \text{const} = 1 \).
which corresponds with rare events with nonlinear $Q$. For $\text{const}_2 \in \{0,1\}$, we generate the missingness indicator in two different ways as follows: $\Delta \sim \text{Bern}(\expit(-1 + A + \text{const}_2 \cdot 2 \cdot \sin(6 \cdot A) + W_1 + (W_2 + W_3)/2))$ (on average 35% missing), $\text{const}_1 = 1$ corresponds with complex outcome missingness, and $\text{const}_2 = 0$ corresponds with simple outcome missingness. Due to confounding bias, we expect the estimator of Donovan et al. to perform poorly. We also expect the binTMLE to be asymptotically biased due to the strong dependence of $\Delta$ on $A$. The results of the simulations are given in Figure 2 and Figure 3. For the complex outcome missingness case, both the Donovan et al. estimator and the binTMLE are biased with especially poor confidence interval coverage as sample size increases. The srTMLE obtains the nominal 95% confidence interval coverage at around sample size $n = 1000$. For very small sample sizes, the bias of the binTMLE is comparable to that of the srTMLE, which is likely due to the sinusoidal signals being indistinguishable from noise and the resulting finite sample bias happening to be favorable for the binTMLE. For the simple outcome missingness case, given in Figure 3, we see both the binTMLE and srTMLE perform similarly in coverage and standard error. Even though the outcome missingness is informed by the biomarker, the binTMLE is only slightly biased and therefore still performs well. This suggests in some cases that the binTMLE, although biased, can still perform well when the outcome missingness is informed by the biomarker.

6.3 Comparison of unadjusted Donovan estimator versus adjusted srTMLE at various levels of confounding

To understand the effect of covariate adjustment in reducing confounding bias, we evaluate the absolute bias $|E_0[\Psi_v^{\text{adj}}(P_{n,v}) - \Psi_v^{\text{adj}}(P_0)]|$ of the Donovan et al. (unadjusted) estimator and (covariate-adjusted) srTMLE at sample sizes $n = 500, 1000, 2000$ for rare event settings with varying levels of confounding. Holding the correlation between $W$ and $A$ fixed at $.5$, we measured the degree of confounding as the amount of correlation between the univariate confounding variable ($W$) and the outcome ($Y$). Across all simulations, we kept the average risk fixed at $0.04$. No outcome missingness was included for simplicity and the simulation design can be found in Supporting Information D. The results given in Figure 4 demonstrate that confounding bias can be significant in rare event settings. Even though the correlation between $A$ and $Y$ is small in such settings, the magnitude of the estimand is also small and thus a small absolute confounding bias can still lead to a large relative bias in the estimates.

7 APPLICATION

We apply the methods developed in this article to the same CYD14 and CYD15 dengue vaccine trial data sets analyzed by Donovan et al. (2019). CYD14 and CYD15
FIGURE 2 Results for second simulation (complex outcome missingness): (A) standard error, (B) absolute bias, (C) mean-squared error, and (D) confidence interval coverage computed from 1000 Monte Carlo simulations for the proposed sTMLE, the binTMLE, and the Donovan-unadjusted estimator. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

FIGURE 3 Results for second simulation (simple outcome missingness): (A) standard error, (B) absolute bias, (C) mean-squared error, and (D) confidence interval coverage computed from 1000 Monte Carlo simulations for the proposed sTMLE, the inefficient binTMLE, and the Donovan-unadjusted estimator. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.
FIGURE 4  Simulation results depicting relative bias (absolute bias divided by estimand value) for the Donovan estimator and covariate-adjusted srTMLE for various levels of confounding correlation (of the confounder $W$ with $Y$) in rare event settings with average risk $\approx 0.04$. Estimates are based on 500 Monte Carlo estimates at sample sizes $n = 500, 1000, 2000$. Cor(W,A) = .5 across all simulations. Uncertainty intervals, defined as $1.95/\sqrt{500}$ times the standard deviation of Monte Carlo estimates, are also displayed. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

were Phase 3 placebo-controlled trials that evaluated the efficacy of the dengue vaccine CYD-TDV in children. CYD14 was conducted in five Asian-Pacific countries with participants aged between 2 and 14 years. CYD15 was conducted in five Latin American countries with participants aged between 9 and 16 years. The two study designs were harmonized allowing for an analysis of the pooled data, where for pooled analysis we restrict to 9- to 16-year-olds given the vaccine is approved for this age range. Doses were given at the start of the study (month 0), month 6, and month 12, with follow-up visits at month 13 and month 25. The primary objective assessed vaccine efficacy against the dengue disease primary endpoint occurring between months 13 and 25 in the per-protocol population, where per-protocol was defined as receiving all three immunizations and not experiencing the dengue endpoint between months 0 and 13. Both studies used a case-cohort sampling design where a simple random sample of participants was selected for measurement of antidengue neutralizing antibody titers at month 13, augmented with these titer measurements for all per-protocol dengue endpoints (Moodie et al., 2018). Using the same convention as Donovan et al., the marker of interest, month 13 log10 neutralizing antibody titer, is defined as the average of the four log10 antibody titers to the four serotypes represented inside the vaccine.

We perform the threshold analysis for both data sets separately and adjust for age, sex, and country. The biomarker variable $A$ is defined as the antibody titer and the outcome $Y$ is defined as $Y = 1(T \leq t_f)$ where $T$ is the time from the month 13 visit to observed dengue endpoint diagnosis and $t_f$ is a reference time point defined as 336 days after the month 13 visit. Since 99.8% of participants were evaluable for whether they experienced the dengue endpoint by month 25, we omitted all individuals censored before time $t_f$ from the analysis at a negligible increase in bias. To estimate the adjusted threshold-response function, we apply the srTMLE defined in Section 5.1.1 with efficient IPW adjustment to account for the cumulative-case control sampling design (see Supporting Information E for how to adjust the TMLE). The IP-weights and all other nuisance functions are estimated nonparametrically.
using the highly adaptive lasso (Benkeser and van der Laan, 2016). We also estimate the unadjusted threshold-response function using the IPW-weighted estimator of Donovan et al., and for comparability, the same IPW weights as the srTMLE estimator were used. The estimated adjusted (TMLE) and unadjusted (Donovan) threshold-response functions with pointwise 95% confidence intervals are given in Figures 5A-C. Figure 6 displays three plots of the reverse-CDF (RCDF) of the immune-response biomarker as a function of the threshold by various covariate strata, and a plot of the estimated expected outcome within levels of the propensity score $P(A \geq \nu \mid W)$ for the pooled CYD14 + CYD15 analysis.

We see that month 13 antibody titer is inversely associated with dengue, where subgroups with threshold value above 3.2 (1585 on natural scale) have on average sevenfold lower estimated risk than those with a value above 1.5 (32 on natural scale) (pooled data). There is little to no difference between the adjusted and unadjusted threshold-response function estimates for the CYD14 study. On the other hand, for the CYD15 analysis and pooled analysis there are some noticeable differences between the adjusted and unadjusted threshold-response function estimates for larger thresholds. For the pooled analysis, the unadjusted estimates are around 20%-30% less than the adjusted estimates for the larger thresholds. Figure 6 (bottom right) shows that there is reasonable variation of the risk of disease acquisition within levels of the estimated propensity score and that the estimated propensity score values evaluated at observations are close to homogeneously distributed on [0.2,0.8]. Figure 6 (top and bottom left) shows that the distribution of the immune-response biomarker are fairly similar within covariate strata. However, the relative difference in the RCDF becomes more significant as the threshold increases, which suggests the presence of more noticeable confounding bias for larger thresholds.
8 | CONCLUSIONS

In this article, we developed a novel nonparametric efficient and double-robust targeted minimum-loss estimator (srTMLE) for the covariate-adjusted threshold-response function with informative outcome missingness, extending the previous work of Donovan et al. (2019). We also presented an inefficient and possibly biased estimator (binTMLE) for the threshold-response that equals the binary treatment-specific mean TMLE (van der Laan and Rose, 2011) where the continuous biomarker is discretized as above or below a threshold. We showed that the threshold-response at a given threshold can be causally identified with a stochastic intervention whose support is restricted to be above the threshold. The theoretical properties of the novel srTMLE were confirmed in a variety of simulation settings, and its performance was compared to the binTMLE and the Donovan et al. unadjusted estimator. Finally, we applied the method to the CYD14 and CYD15 trial data to estimate the covariate-adjusted threshold-response for associating neutralizing antibody titer in vaccine recipients with dengue disease incorporating IPW to accommodate case-cohort sampled antibody titer data. We compared the results with the unadjusted threshold-response estimator as presented in Donovan et al. (2019). The difference between

FIGURE 6  (Top + bottom left) Plots of the reverse-CDF as a function of the biomarker threshold by covariate stratum. (Bottom right) A plot of the estimated expected outcome within levels of the estimated propensity score—$E_0[Y \mid P(A \geq v \mid W)]$ as a function of the estimated propensity score where $v = \text{Median}(A)$. Marginal histograms of the estimates of $E_0[Y \mid P(A \geq v \mid W)]$ and $P(A \geq v \mid W)$ evaluated at observations pooled over the CYD14 and CYD15 studies are also shown. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.
the unadjusted and adjusted estimates was negligible, although some small differences were observed for the trial-pooled analysis. While in this article we considered estimation of \( E_{0,W_0}[Y \mid A \geq v, W] \), the results can be applied to the parameter \( E_{0,W_0}[Y \mid A < v, W] \) by defining \( A := -A \). Various relative and additive contrasts of threshold-response-type estimands can be estimated using a substitution estimator based on the proposed srTMLE and inference can be obtained from the delta method. A useful extension of our work would be toward estimation of the threshold-specific survival function \( E_{0,W}(T > t \mid A \geq v, W) \). While out of the scope of this article, a hazard-based TMLE or sequential regression–based TMLE for this parameter can be developed using similar arguments as in this article. Also, interesting future work would be the investigation of immune marker surrogate endpoints based on the threshold-response function.

**ACKNOWLEDGMENTS**

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number 2R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank the participants of the CYD14 and CYD15 trials and our Sanofi-Pasteur colleagues who conducted these trials. We would like to thank the reviewers and the editor for their helpful comments, which led to numerous improvements to the article.

**DATA AVAILABILITY STATEMENT**

The CYD14 and CYD15 data are available upon request to the sponsor of the studies, Sanofi Pasteur. Data requests can be made through the following url: https://vivli.org.

**ORCID**

Lars van der Laan https://orcid.org/0000-0001-8220-6210

**REFERENCES**

Bang, H. and Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Benkeser, D. and van der Laan M. The highly adaptive lasso estimator, 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016, pp. 689–696.

Bickel, P.J., Klaassen, C.A., Ritov, Y. and Wellner, J. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*, Volume 4. Baltimore, MD: Johns Hopkins University Press.

Cai, W. and van der Laan, M.J. (2020) One-step targeted maximum likelihood estimation for time-to-event outcomes. *Biometrics*, 76, 722–733.

Callegaro, A. and Tibaldi, F. (2019) Assessing correlates of protection in vaccine trials: statistical solutions in the context of high vaccine efficacy. *BMC Medical Research Methodology*, 19(1), 1–7.

Chan, I., Shu, L., Matthews, H., Chan, C., Vessey, R., Sadoff, J. et al. (2002) Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine*, 21, 3411–3430.

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

Díaz, I., Williams, N., Hoffman, K.L. and Schenck, E.J. (2021) Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 1–16. Available from: https://doi.org/10.1080/01621459.2021.1955691

Donovan, K., Hudgens, M. and Gilbert, P.B. (2019) Nonparametric inference for immune response thresholds of risk in vaccine studies. *Annals of Applied Statistics*, 13, 1147–1165.

Genovese, C. and Wasserman, L. (2008) Adaptive confidence bands. *The Annals of Statistics*, 36, 875–905.

Gruber, S. and van der Laan, M.J. (2009) Targeted maximum likelihood estimation: a gentle introduction. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 252.

Moodie, Z., Juraska, M., Huang, Y., Zhuang, Y., Fong, Y., Carpp, L. et al. (2018) Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *Journal of Infectious Diseases*, 217(5), 742–753.

Moore, K.L. and van der Laan, L.M. (2009) Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistical Medicine*, 28(1), 39–64.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*, 2nd edition. London: Cambridge University Press.

Porter, K., Gruber, S., van der Laan, M. and Sekhon, J. (2011) The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, 7(1), 31.

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(427), 846–866.

Rubin, D. (2000) Comment on “Causal inference without counterfactuals,” by A.P. Dawid. *Journal of the American Statistical Association*, 95, 435–437.

Siber, G., Chang, I., Baker, S., Fernsten, P., O’Brien, K., Santosham, M. et al. (2007) Estimating the protective concentration of antimeningococcal capsular polysaccharide antibodies. *Vaccine*, 25, 3816–3826.

Stitelman, O., Hubbard, A. and Jewell, N. (2010) The impact of coarsening the explanatory variable of interest in making causal inferences: implicit assumptions behind dichotomizing variables. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 264.

Taubman, S.L., Robins, J.M., Mittleman, M.A. and Hernán, M.A. (2009) Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6), 1599–1611.

Vaart, A.W.v.d. (1998) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

van der Laan, M.J. and Rose, S. (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

van der Laan, M.J. and Robins, J.M. (2003) *Unified methods for censored longitudinal data and causality*, Volume 5. New York: Springer.
SUPPORTING INFORMATION
Web Appendices referenced in Sections 2.2, 2.3, 3.1, 3.2, 3.3, 4.1 are available with this paper at the Biometrics website on Wiley Online Library: (A) Derivation of efficient influence function of target parameter and causal identification. (B) Proof of efficiency of the TML estimators. (C) Definition and discussion of Donovan estimator. (D) Simulation designs. (E) Adjusting TMLE when biomarker or treatment is missing at random. (F) Some miscellania regarding inference, testing, causal interpretation. (G) Nuisance parameter estimation and computational considerations. (H) R code implementing the estimators given in Section 5. (I) Applying method to bounded continuous outcomes.

How to cite this article: van der Laan L., Zhang W., Gilbert P.B. (2023) Nonparametric estimation of the causal effect of a stochastic threshold-based intervention. Biometrics, 79, 1014–1028. https://doi.org/10.1111/biom.13690