Estimating optimal surrogate endpoints by machine learning and targeted minimum loss-based estimation in two-phase sampling studies

Brenda Price

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Peter Gilbert, Chair

Alex Luedtke

Marco Carone

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Estimating optimal surrogate endpoints by machine learning and targeted minimum loss-based estimation in two-phase sampling studies

Brenda Price

Chair of the Supervisory Committee:
Dr. Peter Gilbert
Department of Biostatistics

This dissertation provides contributions in two areas: the application of TMLE in estimation of an optimal surrogate and implementation of inverse probability of censoring weighted targeted minimum loss-based estimation (IPCW-TMLE). In Chapter 1 we develop methodology for the estimation of optimal surrogates in randomized trials using targeted minimum loss-based estimation (TMLE), first in the setting of complete data, and then in Chapter 2, extended to the setting of two-phase data, seeking to make the methodology more applicable to real randomized trials. In Chapter 3 we present a comparison of IPCW-TMLE to a commonly used method of Breslow and Holubkov for parameter estimation in two-phase studies. The simulation study presented assesses the comparative differences in bias and efficiency of estimates obtained by both methods. In Chapter 4, IPCW-TMLE is elaborated for estimation of causal parameters of interest in right-censored two-phase studies. The methods developed in this dissertation have broad application to randomized clinical trials with two-phase designs for measuring biomarkers. Many of the methods described in this dissertation are illustrated with application to two dengue phase 3 vaccine efficacy trials.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to her advisor, Peter Gilbert for his continued support and guidance, and to Mark van der Laan for his collaboration on the estimated optimal surrogate. She appreciates the time and effort contributed by all the members of her committee: Marco Carone, Alexander Luedtke, Susanne May, Linda Shapiro, and Su-In Lee. She would also like to thank the CYD14 and CYD15 study participants, her many collaborators at Sanofi Pasteur, the protocol leadership, and the numerous groups and individuals dedicated to planning and conducting these two trials, as well as the HVTN 505 study participants, protocol leadership, and the many groups and individuals dedicated to planning and conducting the HVTN 505 trial, with special thanks to Georgia Tomaras and Julie McElrath for their labs' generation of the immunology data analyzed in the HIV vaccine data application.

DEDICATION

to Nathan,

Camille, Madeline,

and Sophia

Chapter 1

ESTIMATION OF THE OPTIMAL SURROGATE BASED ON A RANDOMIZED TRIAL

1.1 Introduction

A common scientific problem is to determine a surrogate outcome for a long-term outcome so that future randomized studies can restrict themselves to only collecting the surrogate outcome. We consider a study where we observe n independent and identically distributed observations of a random variable consisting of baseline covariates, a treatment, a vector of candidate surrogate outcomes measured at or before the intermediate time point, and the outcome of interest at a final time point. We assume that the treatment is randomized, conditional on the baseline covariates. The goal is to use these data to produce a candidate surrogate that is maximally promising for use in future trials for estimation and testing of a mean contrast treatment effect on the final outcome. We define an optimal surrogate for the current trial as the function of the data collected by the intermediate time point that optimally predicts the final outcome: this optimal surrogate is a function of the data generating distribution and is thus unknown. In Section 1.2 we show the desirable property of this optimal surrogate that the subgroup-specific average causal effect of treatment on the optimal surrogate is identical to the subgroup-specific average causal effect of treatment on the final outcome. It follows that, by construction, this optimal surrogate satisfies the Prentice [50] definition of a valid surrogate endpoint for all subgroups defined by the baseline covariates. An important property of a valid Prentice surrogate is that the disastrous 'surrogate para-

dox' [defined as (i) the effect of the treatment on the surrogate is positive, (ii) the surrogate and outcome are strongly positively correlated, but (iii) the effect on the treatment on the outcome is negative] cannot occur [68]. In addition, the average causal effect on the optimal surrogate has the same interpretation as the average causal effect on the clinical endpoint, such that, appealingly, the surrogate effect has the same interpretation as the clinical effect.

In Section 1.3 we show that the optimality of the surrogate (and thus its Prentice-validity) is invariant to changes in the joint distribution of the covariates, treatment, and intermediate outcomes, as long as treatment remains conditionally randomized given the baseline covariates in the new trial and the conditional mean of the outcome is unchanged. Thus under these conditions, in such a new trial the average treatment effect on the optimal surrogate (optimized in the current trial and applied in the new trial) equals the average treatment effect on the final outcome. This implies that, in a thought experiment where the current trial has an infinite sample size such that the optimal surrogate itself is measurable and is used as the surrogate in the new trial, a $(1-\alpha)\%$ confidence interval for the optimal surrogate treatment effect parameter is also a $(1-\alpha)\%$ confidence interval for the clinical treatment effect parameter. Finally, if, in addition, the final outcome only depends on the treatment through a vector of candidate surrogate outcomes in both the current and the new study [i.e., Prentice's 'full mediation' criterion holds [50]], then the optimal surrogate does not depend on treatment, so that it is also the optimal surrogate in the new study, even when it involves a completely different treatment.

In Section 1.4 we present a super-learner estimator of the optimal surrogate, thereby incorporating the state of the art in machine learning and nonparametric estimation in an asymptotically optimal way. The cross-validated mean squared error can be used as an objective measure of performance of the surrogate in predicting the final outcome, and the literature also provides a confidence interval for the true mean squared error of the super-learner estimator when applied to the training samples in the cross-validation scheme. The

latter is an important target parameter itself since it measures the strength of the surrogate. In Section 1.5 we further propose to target the super-learner fit of the optimal surrogate with targeted minimum loss-based estimation (TMLE), such that the estimator of the effect of treatment on this targeted estimated optimal surrogate is an asymptotically linear and efficient estimator of the average causal effect of treatment on the outcome of interest in the current trial. [The targeting optimizes the bias-variance tradeoff specifically for estimating the target parameter of interest, a general feature of TMLE as discussed in van der Laan and Rose [65].] Whereas the TMLE is the most advantageous estimator of the optimal surrogate, adding the targeting step does not improve the ability to generalize inferences to new settings, such that the super-learner alone is a sound strategy for generating promising candidate surrogate endpoints.

Our objective is to develop a most-promising surrogate outcome based on a clinical outcome study with possibly high-dimensional candidate surrogates; in future work we plan to address the related important objective of using the developed surrogate outcome as an endpoint in a future study to make inference (i.e., construct confidence intervals) on the causal effect of treatment in that setting without measuring the clinical outcome (future work is needed because inference based on nonparametric super-learning is a hard problem). However, in Section 1.3.4 we discuss approaches to inference for the future study based on the previously developed estimated optimal surrogate, accounting for the estimation error. We stress that because the assumptions needed for bridging clinical efficacy based on a surrogate endpoint to a new setting (stated in theorems 2 and 3) are generally difficult to verify, it is recommended that wherever possible (e.g., not prohibited by ethics) future efficacy trials assess efficacy directly based on the true clinical endpoint. In Section 1.6 we apply the proposed approach to two simulated trials based on two recent dengue vaccine efficacy trials, building the estimated optimal surrogate from one trial and using it for estimating the clinical treatment effect in the second trial. In Section 1.7 we conduct a simulation study of the

proposed methodology, and in Section 1.8 conclude with remarks.

1.1.1 Connection of the optimal surrogate framework to other surrogate frameworks

The newly proposed framework does not fit squarely into any of five existing frameworks for surrogate endpoints– the Prentice [50] replacement endpoint framework, the controlled direct and indirect causal effects framework [34, 53] the principal stratification framework [18], the meta-analysis framework [13, 8], and the causal selection diagram framework [46]. It is more similar to the Prentice, meta-analysis, and causal selection diagram frameworks, in being based purely on statistical parameters that are estimable under the basic assumptions typically made in randomized clinical trials. In particular, it aligns most closely with the Prentice framework by taking as its starting point the Prentice definition of a valid surrogate endpoint. In fact, the optimal surrogate is constructed to guarantee satisfaction of the Prentice definition, a unique advantage compared to previous approaches. Our approach also departs from previous approaches by defining the optimal surrogate as an unknown parameter, such that its predicted values– which are the realized values of the estimated optimal surrogate– are used as the surrogate endpoint. Under standard assumptions of randomized trials, if the estimated optimal surrogate is consistent for the optimal surrogate as attained via nonparametric learning, then for large sample size trials it must approximately satisfy the Prentice definition. Section 1.9 elaborates the connections of the optimal surrogate framework with the other surrogate frameworks.

The optimal surrogate approach breaks new ground by treating the surrogate endpoint problem as a supervised statistical learning problem. While historically methods evaluate a pre-selected univariable or low-dimensional vector candidate surrogate, the optimal surrogate approach allows all collected baseline and intermediate response data to potentially contribute to the optimal surrogate, selected and combined through unbiased machine learning, and does not require parametric modeling assumptions.

1.2 Statistical Formulation of Estimation of an Optimal Surrogate

Suppose we have observed n independent and identically distributed copies of a time-ordered data structure $O = (W, A, S, Y) \sim P_0$, where W is a vector of baseline covariates, A is a binary treatment assigned at baseline, and S is a vector of intermediate outcomes measured at (or before) some time point τ, and Y is the final univariate outcome of interest measured at a final time point after τ. We assume a structural causal model [65] $W = f_W(U_W)$, $A = f_A(W, U_A)$, $S = f_S(W, A, U_S)$, $Y = f_Y(W, A, S, U_Y)$ defined by a collection of deterministic functions f_W, f_A, f_S, f_Y and a probability distribution P_U for the exogenous errors $U = (U_W, U_A, U_S, U_Y)$, where the structural model makes no assumptions on f_W, f_S, f_Y, but possibly some assumptions on f_A (e.g., in a randomized controlled trial, f_A is known). Let $S_1 = f_S(W, A = 1, U_S)$ and $S_0 = f_S(W, A = 0, U_S)$ denote counterfactual surrogates under treatment and control, respectively, and similarly let $Y_1 = f_Y(W, A = 1, S_1, U_Y), Y_0 = f_Y(W, A = 0, S_0, U_Y)$ denote the counterfactual outcomes. The structural causal model also assumes that the treatment is randomized conditional on W: U_A is independent of (U_S, U_Y), given W.

Let $X = (W, S_0, S_1, Y_0, Y_1)$ denote the full-data structure of interest, with probability distribution $P_{X,0}$. The observed data distribution P_0 of O is determined by the full-data distribution $P_{X,0}$ and the conditional distribution g_0 of A, given X, where $g_0(a \mid X) = g_0(a \mid W)$. The statistical model for P_0 makes at most some assumptions about the conditional distribution g_0 of A given W. For example, if it is a randomized trial, then g_0 is known. Thus the statistical model $\mathcal{M}$ for P_0 only (possibly) constrains g_0, but puts no assumptions on the marginal distribution of W nor on the conditional distribution of (S, Y), given A, W. In this setting we also make the assumption of positivity: $P_0(A = a|W) > 0$ a.e. for $a \in \{0, 1\}$.

In future studies one hopes to replace the final outcome Y by a so-called surrogate outcome measured by the intermediate time point τ. We refer to any real-valued function $(W, A, S) \to \psi(W, A, S) \in \mathbb{R}$ as a candidate surrogate, representing a measurement one can

collect by time τ. If one wants to consider surrogates that depend on S only through a subset/summary of the S, then the setting is simply applied to S defined by this subset. In addition, if it is known that $g_0(A \mid W)$ only depends on W through W_1, and one wants to only consider surrogates that depend on W through W_1, then the setting is applied to $W = W_1$. These restrictions will come at a price for the external validity of the proposed surrogate, as will become clear, so that these choices should not be taken lightly.

The key question is now how are we going to define a good surrogate, defined in terms of the true data distribution P_0? To start with we want the surrogate outcome $S^\psi \equiv \psi(W, A, S)$ to be a valid surrogate in the actual study, according to the Prentice definition: that is, $E_0(Y_1 - Y_0) = 0$ if and only if $E_0(S_1^\psi - S_0^\psi) = 0$, where the counterfactual $S_a^\psi = \psi(W, a, S_a)$, $a \in \{0, 1\}$. This guarantees that in this particular study involving sampling from P_0, a test for $H_0^\psi : E_0(S_1^\psi - S_0^\psi) = 0$, which controls the type-I error at level α, yields a test for $H_0 : E_0(Y_1 - Y_0) = 0$ with type-I error control at level α, where the latter test is simply defined by rejecting H_0 if and only if H_0^ψ is rejected. Importantly, by estimating $E_0(Y_1)$ and $E_0(Y_0)$ separately, our approach applies for a general treatment effect contrast.

We also need a criterion depending on P_0 that can be used to rank valid surrogates based on the data $O_1, \ldots, O_n$, and to define a P_0- optimal surrogate with respect to that criterion. In this manner, we not only select a P_0-valid surrogate but a P_0-optimal one in the class of P_0-valid surrogates. We would like to select the criterion such that the P_0-optimal surrogate is not only optimal under P_0 with respect to this criterion, but that being P_0-optimal implies that the validity of the optimal surrogate is invariant to a variety of possible changes in the data generating experiment. Or, even better, we would like that the P_0-optimal surrogate is also a P-optimal surrogate (and thus valid) under a variety of P's different from P_0.

For these purposes, our proposed criterion is the following full-data mean squared error:

$$\psi \to MSE_{P_{X,0}}(\psi) \equiv \sum_a E_{P_{X,0}} \left\{ g_0(a \mid W)(Y_a - \psi(W, a, S_a))^2 \right\}. \tag{1.1}$$

That is, our goal is to minimize the weighted mean square prediction error for predicting the actual counterfactual outcome of interest, across the different treatment values, with constraint that the solution must satisfy the Prentice definition as stated above. The idea is that if a subject is assigned treatment $A = a$ and one uses as surrogate outcome $S_a^\psi = \psi(W, a, S_a)$, then one wants that surrogate outcome to be a good approximation of the future outcome Y_a. Depending on the future use of the surrogate, one might put more weight on doing well in predicting Y_1 for treated subjects versus predicting Y_0 for untreated subjects, in which case this particular weighting scheme $g_0(a \mid W)$ needs to be replaced by another weighting scheme. Given a class $\mathbf{\Psi}$ of possible surrogate functions $\psi()$ satisfying the Prentice definition by construction, the P_0-optimal surrogate in this class is then defined as

$$\psi_0^F = \arg\min_{\psi \in \mathbf{\Psi}} MSE_{P_{X,0}}(\psi).$$

In this article, we focus on the nonparametric class $\mathbf{\Psi}$ consisting of all functions of (W, A, S). In that case, the choice of weight in $MSE_{P_{X,0}}$ (i.e., $g_0(a \mid W)$) does not affect the optimal solution: i.e., the optimal surrogate will be optimal for each choice of weight. The P_0-optimal surrogate ψ_0^F is given by

$$\psi_0^F(w, a, s) = E_0(Y_a \mid W = w, S_a = s),$$

which is a standard solution to a minimization problem that is the same whether or not the Prentice definition constraint is used. Since A is conditionally independent of (U_S, U_Y), given W, it follows that the full-data MSE equals the observed data MSE:

$$MSE_{P_{X,0}}(\psi) = MSE_{P_0}(\psi) \equiv E_{P_0}(Y - \psi(W, A, S))^2.$$

As a consequence, under our assumption that A is randomized, ψ_0^F is identifiable from P_0

and can also be defined as:

$$\psi_0^F(W, A, S) = \psi_0(W, A, S) \equiv E_0(Y \mid W, A, S).$$

In other words, due to the randomization of A, we have $E_0(Y_a \mid W = w, S_a = s) = E_0(Y \mid W = w, S = s, A = a)$. It also follows that $E_{P_0}(\psi_0(W, a, S_a) \mid W) = E_{P_0}(Y_a \mid W)$, which demonstrates that the treatment specific counterfactual mean of the P_0-optimal surrogate equals the treatment specific counterfactual mean of the outcome. This shows that an average causal effect of treatment on the P_0-optimal surrogate equals the desired average causal effect of treatment on the outcome. Thus, under P_0, an $(1-\alpha)\%$ confidence interval for the causal effect of treatment on the P_0-optimal surrogate is also a $(1-\alpha)\%$ confidence interval for the causal effect of treatment on the outcome Y. We state this as a theorem.

THEOREM 1: *Assume positivity:* $P_0(A = a|W) > 0$ *a.e. for* $a \in \{0, 1\}$. *Then the minimizer of the counterfactual mean squared error* $\psi \to MSE_{P_{X,0}}(\psi)$ *over all functions* $(W, A, S) \to \psi(W, A, S)$ *satisfying the Prentice definition of a valid surrogate endpoint is given by:*

$$\bar{S}_0 = \psi_0(W, A, S) \equiv E_0(Y \mid W, A, S).$$

We call this the P_0*-optimal surrogate. We also note that the counterfactuals of this* P_0*-optimal surrogate are given by:* $\bar{S}_{0,a} = E_0(Y_a \mid W, S_a)$, $a \in \{0, 1\}$. *In addition,*

$$E_{P_0}(\bar{S}_{0,a} \mid W) = E_{P_0}(Y_a \mid W).$$

This shows that the P_0-optimal surrogate has the desirable properties of a valid surrogate in the actual P_0-study. Moreover, if each treatment $A = a$ is considered separately, then the minimizer of $\psi_a \to MSE_{P_{X,a,0}}(\psi_a)$ over all functions $(W, a, S) \to \psi_a(W, a, S)$ is $E_0(Y|W, A = a, S)$, where $MSE_{P_{X,a,0}}$ is the a^{th} term in the sum $MSE_{P_{X,0}}(\psi)$ in (2.3).

Therefore the P_0-optimal surrogate is the same whether one minimizes the overall MSE in (2.3) or minimizes the treatment-specific MSEs separately (as we do in the simulations and application). Theorem 1 motivates us to define the P_0-optimal surrogate as the desired surrogate, and we will further support this by establishing that the P_0-optimal surrogate is also a P-optimal surrogate under a variety of distributions P different from P_0.

The statistical estimation problem for the original trial is now defined: we observe n i.i.d. $O \sim P_0 \in \mathcal{M}$, the target parameter mapping is defined by $\Psi : \mathcal{M} \to \mathbf{\Psi}$ with $\Psi(P) = E_P(Y \mid W, A, S)$, and $\psi_0 = E_{P_0}(Y \mid W, A, S)$ is the true value we aim to learn from the data.

1.3 Conditions on the New Study P Under Which the P_0-Optimal Surrogate is Also the P-Optimal Surrogate

1.3.1 Invariance of the P_0-optimal surrogate to changes in the distribution of (W, A, S)

The following theorem is a trivial consequence of the fact that $E_{P_0}(Y \mid W, A, S)$ does not depend on the choice of joint distribution of (W, A, S), and $E_P(Y \mid W, A = a, S = s) = E_P(Y_a \mid W, S_a = s)$ if A is randomized in the P-world. Nonetheless, it demonstrates that the P_0-optimal surrogate is also the P-optimal surrogate in any study P that only differs in the joint distribution of (W, A, S), and preserves the conditional randomization of treatment. We assume both the current and future studies are randomized studies for data structures (W, A, S, Y) and (W^*, A^*, S^*, Y^*) with probability distribution P_0 and P, respectively; for the current study this means that we have the structural equation model $W = f_W$, $A = f_A$, $S = f_S$, $Y = f_Y$ and U_A is independent of (U_S, U_Y), given W.

THEOREM 2: *Assume (1) the current and future randomized studies defined above satisfy Equal Conditional Means: $E_P(Y^*|W^* = w, A^* = a, S^* = s) = E_{P_0}(Y|W = w, A = a, S = s)$ for all (w, a, s) in a support of (W^*, A^*, S^*), (2) a support of (W^*, A^*, S^*) is contained in a support of (W, A, S), and (3) positivity: $P_0(A = a|W) > 0$ a.e. and $P(A^* = a|W^*) > 0$ a.e. for $a \in \{0, 1\}$. Then, the P_0-optimal surrogate equals the P-optimal surrogate.*

Thus, Equal Conditional Means, contained support, and positivity are sufficient conditions to make the P_0-optimal surrogate still a valid surrogate in a new randomized study that differs in the marginal distribution of W, in the conditional distribution of A given W, and in the conditional distribution of S given A, W.

1.3.2 Generalizability when the surrogate completely blocks the effects of both treatments

Normally, if the new study considers a whole different treatment than in the current study, then its effect on the outcome will be different and one would thus expect that the conditional mean of Y, given W, A, S, will be modified as well. Therefore, the conditions on the new study P in the previous theorem essentially exclude studies that evaluate a new treatment. However, there is one very important exception. The following theorem is just a special case of the previous theorem, but the implication is that if $f_Y(W, 0, S_a, U_Y) = f_Y(W, 1, S_a, U_Y)$ for $a \in \{0, 1\}$ such that the outcome Y only depends on the treatment through its effect on the surrogate vector (i.e., Prentice's 'full mediation' criterion), then the new study can even consider a whole different treatment as long as it also only affects Y through S again. That is, if S is rich enough that it blocks the effect of any future treatment on the outcome, then our P_0-optimal surrogate can also be used in future studies evaluating different treatments. The key behind this result is that under this no-direct effect assumption the P_0-optimal surrogate $E_{P_0}(Y \mid W = w, A = a, S = s)$ at (w, a, s) is constant in a and is thus only a function of (w, s).

THEOREM 3: *In addition to the conditions of Theorem 2, assume* $E_0(Y|W, A, S) = E_0(Y|W, S)$. *Then, the* P*-optimal surrogate equals the* P_0*-optimal surrogate and* $E_P(Y^* \mid W^* = w, A^* = a, S^* = s) = E_P(Y^*_a \mid W^* = w, S^*_a = s)$. *In addition,* $a \mapsto E_{P_0}(Y_a \mid W = w, S_a = s)$ *and* $a \mapsto E_P(Y^*_a \mid W^* = w, S^*_a = s)$ *are constant in* a.

The first statement of Theorem 3 is established by Theorem 2, so that we only need to show the last statement. By assumption $a \to E_{P_0}(Y \mid W = w, A = a, S = s)$ is constant in

a for P_0-a.e (w, s). Thus, $a \mapsto E_{P_0}(Y_a \mid W = w, S_a = s)$ is constant in a for P_0-a.e. (w, s), but since $E_P(Y^* \mid W^* = w, A^* = a, S^* = s) = E_{P_0}(Y \mid W = w, A = a, S = s)$ for all P-a.e. (w, s) (since the support of (W^*, S^*) is contained in the support of (W, S)), we also have that $a \mapsto E_P(Y^* \mid W^* = w, A^* = a, S^* = s)$ is constant, and, by randomization of A^*, the latter is equivalent to $a \mapsto E_P(Y^*_a \mid W^* = w, S^*_a = s)$ is constant. □

1.3.3 How to define the surrogate in a future study when the transportability assumptions fail?

When the new study evaluates a new treatment A^*, typically it is not reasonable to assume that the intermediate variables S completely block the effect of treatment (current and new) on the outcome. Section 1.3.4 discusses how $E_{P_0}(Y \mid W, A, S)$ may still often be a good candidate surrogate for such a future study, and discusses implications about differences between $E_P(Y^* \mid W^* = w, A^* = a, S^* = s)$ and $E_{P_0}(Y \mid W = w, A = a, S = s)$ for $a \in \{0, 1\}$.

1.3.4 Inference on the clinical treatment effect in a future study based on the previously estimated optimal surrogate, accounting for estimation error and failure of the transportability assumptions

In Sections 1.3.1 and 1.3.2 we showed conditions on the new study P and current study P_0 under which the P_0-optimal surrogate is also the P-optimal surrogate. That is the best possible scenario as statistical inference for the causal effect of treatment on the P_0-optimal surrogate outcome in the new trial corresponds exactly with statistical inference for the causal effect of treatment on the outcome of interest in the new trial. We now consider a situation in which the new study evaluates a new treatment A^* and we are not willing to assume that the intermediate variables S completely block the effect of treatment (current and new) on the outcome; this situation will be very common. Now we can be certain that the P_0-optimal

surrogate $E_{P_0}(Y \mid W, A, S)$ is not equal to the P-optimal surrogate, and, since the P_0-optimal surrogate is a function of A which is not measured/evaluated in the new study, one needs to decide how to even define a surrogate for the future study based on $E_{P_0}(Y \mid W, A, S)$. One can imagine that one would use $E_{P_0}(Y \mid W, A, S)$ as a surrogate if we feel that the treatment $A = 1$ in the P_0-study is most comparable with the treatment $A^* = 1$ in the new P-study. Even though we now have no guarantees, $E_{P_0}(Y \mid W, A, S)$ will often be a good candidate surrogate for such a future study (i.e., one that may approximately satisfy the Prentice definition of a valid surrogate in the future P-study), but one needs to be concerned about the difference between $E_P(Y^* \mid W^* = w, A^* = a, S^* = s)$ and $E_{P_0}(Y \mid W = w, A = a, S = s)$ for $a \in \{0, 1\}$.

To address this issue, suppose that ψ_n converges to ψ_0 at a rate $r(n)$ in the sense that $d_0(\psi_n, \psi_0) = O_P(r(n))$, where $d_{P_0}(\psi, \psi_0) = P_0 L(\psi) - P_0 L(\psi_0)$ is the loss-based dissimilarity. We also have that $d_0(\psi_n, \psi_0) = O_P(r(n))$, but for simplicity we work with ψ_n. For concreteness, let us consider the squared error loss $L(\psi)(O) = (Y - \psi(W, A, S))^2$. Consider a new study P with data structure (W^*, A^*, S^*, Y^*) for which the Equal Conditional Means condition holds, and for which we only collect the surrogate $\psi_n(W^*, A^*, S^*)$ instead of Y^*.

In this new study the data structure is $(W^*, A^*, S^*, \psi_n(W^*, A^*, S^*))$ and one would target the parameter $\theta^*_P = \theta^*_\psi(P) = E_P\left[E_P(\psi_n(W^*, 1, S^*) \mid A^* = 1, W^*)\right] - E_P\left[E_P(\psi_n(W^*, 0, S^*) \mid A^* = 0, W^*)\right]$. The clinical treatment effect target parameter of this study is $E_P(Y_1^* - Y_0^*) = \theta^*_P = \theta^*_\psi(P)$

$$E_P\left[E_P(\psi_0(W^*, 1, S^*) \mid A^* = 1, W^*)\right] - E_P\left[E_P(\psi_0(W^*, 0, S^*) \mid A^* = 0, W^*)\right].$$

Suppose that $dP(W^* = w, A^* = a, S^* = s)/dP_0(W = w, A = a, S = s) < M < \infty$ P-a.e. for (w, a, s) in a support of (W^*, A^*, S^*). In that case, it follows that

$$d_P(\psi_n, \psi^*) = \int (\psi_n - \psi^*)^2(W^*, A^*, S^*) dP(W^*, A^*, S^*) \leq M d_{P_0}(\psi_n, \psi^*)$$

where $Md_{P_0}(\psi_n, \psi^*) = O_P(r(n))$. Thus, under this condition, $\theta_{\psi_n}(P) - \theta^*_\psi(P) = O_P(r(n))$.

From this we learn that the estimand defined by the average causal effect of treatment on the surrogate $\psi_n(W^*, A^*, S^*)$ in the future study P will be within distance $O_P(r(n))$ from the desired average causal effect of treatment on the actual outcome Y^*. Suppose that one is only interested in picking up causal effects on Y^* that are larger than some minimal value δ^*. Then, one would want to make sure this remainder $O_P(r(n)) < \delta^*$ so that $\mid \theta_{\psi_n}(P) - \theta^*_\psi(P) \mid < \delta^*$. The difference $\theta_{\psi_n}(P) - \theta^*_\psi(P)$ equals $[\theta_{\psi_n}(P) - \theta_{\psi_0}(P_0)] + [\theta_{\psi_0}(P_0) - \theta^*_\psi(P)]$, showing that the first source of the $O_P(r(n))$ remainder is the discrepancy between the estimated optimal surrogate and the true optimal surrogate in the original trial, and the second source is any violations of the Equal Conditional Means condition. Therefore, under this condition, if the original study were very large such that the first discrepancy is negligible, then the surrogate parameter $\theta_{\psi_\infty}(P)$ studied in the new trial equals the target parameter of interest $\theta^*_\psi(P)$. Thus an infinite original study plus Equal Conditional Means implies that point and confidence interval estimates for $\theta^*_\psi(P)$ can be obtained simply by point and confidence interval estimates for the surrogate effect $\theta_{\psi_\infty}(P)$. In addition, for a finite-sample original study, under Equal Conditional Means $[\theta_{\psi_n}(P) - \theta_{\psi_0}(P_0)]$ measures the bias for estimating $\theta^*_\psi(P)$ based on the estimated optimal surrogate instead of on Y^*. Clearly, the idea is that the estimated optimal surrogate must be a good estimate of the true optimal surrogate in the original study, and $E_{P_0}(Y \mid W = w, A = a, S = s)$ must be a reasonable approximation of $E_P(Y^* \mid W^* = w, A^* = a, S^* = s)$ in the future study, in order to trust our surrogate outcome as a surrogate for the outcome in a future study.

Future work is needed to obtain confidence intervals for $\theta^*_\psi(P) = E_P(Y^*_1 - Y^*_0)$ based on the estimated optimal surrogate instead of on the true optimal surrogate. This problem is readily solved if ψ_n were estimated using a parametric model, in which case the delta method would yield a confidence interval for ψ_0 and for $\theta^*_\psi(P)$, and this parametric model could be selected data-adaptively. However, obtaining a confidence interval when estimating ψ_n

nonparametrically through super-learning as we do is much harder, because ψ_0 is a function that is not estimable at root-n rate.

1.4 Super-learning of the P_0-Optimal Surrogate

Estimation of the P_0-optimal surrogate is a standard prediction problem. That is, we estimate $E_0(Y \mid W, A, S)$ with a minimizer of the risk of a loss: $\psi_0 = \arg\min_\psi P_0 L(\psi)$, with $Pf \equiv \int f(o)dP(o)$. For example, one could use squared error loss $L(\psi)(O) = (Y - \psi(W, A, S))^2$. To construct an optimal estimator among any given class of candidate estimators, we use loss-based super-learning. The oracle inequality for the cross-validation selector guarantees that the estimator is asymptotically at least as good as any candidate in the set of candidate estimators [63, 65].

Let $\hat{\Psi}_j : \mathcal{M}_{NP} \to \Psi(\mathcal{M})$ be a candidate estimator that maps an empirical distribution of $(O_1, \ldots, O_n)$ (i.e., an element of the nonparametric model $\mathcal{M}_{NP}$ of probability distributions) into the parameter space $\Psi(\mathcal{M}) = \{\Psi(P) : P \in \mathcal{M}\}$, $j = 1, \ldots, J$. This library of candidate estimators could include a variety of parametric model based estimators as well as a variety of machine learning algorithms, possibly coupled with different dimension reduction strategies, and possibly indexed by a variety of tuning parameters.

Let $B_n \in \{0, 1\}^n$ be a random split of the sample into a training sample $\{i : B_n(i) = 0\}$ and validation sample $\{i : B_n(i) = 1\}$. For example, if we use V-fold cross-validation defined by a partitioning of the sample in V equal size groups, then B_n has V possible realizations, each occurring with probability $1/V$, and each split corresponds with setting the components of B_n in one of the V-folds equal to 1 and setting the other components equal to 0. Let P^0_{n,B_n} and P^1_{n,B_n} be the empirical distributions of the training and validation sample corresponding with split-vector B_n, respectively. The cross-validated risk of the j-th candidate estimator is then defined as $E_{B_n} P^1_{n,B_n} L(\hat{\Psi}_j(P^0_{n,B_n}))$, where $L(\cdot)$ should be chosen as squared error loss to be consistent with our proposed criterion (1) from Section 1.2.

One could now define the cross-validation selector

$$J_n = \arg\min_j E_{B_n} P^1_{n,B_n} L(\hat{\Psi}_j(P^0_{n,B_n}))$$

as the selector of the winner, and the corresponding discrete super-learner is then defined as $\hat{\Psi}(P_n) = \hat{\Psi}_{J_n}(P_n)$. One could also propose a parametric family $\{f_\alpha : \alpha\}$ of functions from $\mathbb{R}^J$ to the real line that represents a family of combinations of all the J estimators:

$$\hat{\Psi}_\alpha(P_n) = f_\alpha(\hat{\Psi}_j(P_n) : j = 1, \ldots, J),$$

and where α represents a multivariate vector. For example, one might define $\hat{\Psi}_\alpha = \sum_{j=1}^J \alpha_j \hat{\Psi}_j$ as a weighted linear combination of the candidate estimators, where the weights α_j are restricted to be non-negative and sum to 1. One now defines the cross-validation selector for this continuous family of candidate estimators $\{\hat{\Psi}_\alpha : \alpha\}$ accordingly:

$$\alpha_n = \arg\min_\alpha E_{B_n} P^1_{n,B_n} L(\hat{\Psi}_\alpha(P^0_{n,B_n})).$$

The super-learner is then defined as $\hat{\Psi}(P_n) = \hat{\Psi}_{\alpha_n}(P_n)$. By the oracle inequality for the cross-validation selector, the super-learner is asymptotically equivalent with the oracle selected estimator, as long as the realistic assumption holds that none of the candidate estimators is a correctly specified parametric model [63].

In addition, we can evaluate the super-learner by its cross-validated risk, using a cross-validation scheme S_n (e.g., using V-fold cross-validation again as in the super-learner):

$$\text{CV-RISK} = E_{S_n} P^1_{n,S_n} L(\hat{\Psi}(P^0_{n,S_n})), \tag{1.2}$$

which involves rerunning the super-learner on learning samples $\{i : S_n(i) = 0\}$ and evaluating it on test samples $\{i : S_n(i) = 1\}$, and averaging the performance across the different splits.

This represents an estimator of the true conditional risk

$$E_{S_n} R(\hat{\Psi}(P^0_{n,S_n}) \mid P_0) \equiv E_{S_n} P_0 L(\hat{\Psi}(P^0_{n,S_n})),$$

and one can also construct a Wald-type 95% confidence interval for the latter parameter $E_{S_n} R(\hat{\Psi}(P^0_{n,S_n}) \mid P_0)$ given by CV-RISK $\pm\, 1.96\sigma_n/\sqrt{n}$, where $\sigma_n^2 = E_{S_n} P^1_{n,S_n} \left\{ L(\hat{\Psi}(P^0_{n,S_n})) - E_{S_n} P^1_{n,S_n} L(\hat{\Psi}(P^0_{n,S_n})) \right\}^2$. The theory behind the asymptotic correctness of this data adaptive confidence interval is given in van der Laan, Hubbard, and Pajouh [62]. A super-learner can be built and fitted with the R package *superlearner* available at CRAN.

One can also define a cross-validated R^2:

$$\text{CV-R}^2 = 1 - \text{CV-RISK}/E_{S_n} P^1_{n,S_n} L(\hat{\Psi}^0(P^0_{n,S_n})), \tag{1.3}$$

where $\hat{\Psi}^0(P_n) = \int y dP_n(y)$ is the empirical mean of the Y_i-values. This provides a universal measure of the strength of the estimated surrogate $\hat{\Psi}$, allowing us to compare different candidate surrogate estimators across studies and within a study. For example, one might construct a super-learner $\hat{\Psi}_\delta$ based on δ-specific subsets (W_δ, S_δ) of the complete (W, S), where δ is a measure of the complexity of the resulting surrogate as a function of (W, S). One could now plot CV-R^2 of $\hat{\Psi}_\delta$ against δ for a sequence of δ-values, and the user can decide on a choice of δ taking into account both complexity and strength of the surrogate. This analysis is practically important given that all of the variables (W_δ, S_δ) used in the estimated optimal surrogate need to be collected in a future trial to use the estimated optimal surrogate in that trial; in practice some variable sets may be selected based on their high likelihood of being collected.

1.5 The Targeted Estimated Optimal Surrogate Captures All Information About Outcome for the Sake of Estimation of the Average Treatment Effect

One could estimate the optimal surrogate $E_0(Y \mid W, A, S)$ based on any model for the conditional mean. If (W, S) is moderate-to-high dimensional, then it is typically infeasible to attain a consistent estimator of $E_0(Y \mid W, A, S)$ based on a particular parametric model, because of insufficient knowledge. Accordingly we recommend the super-learner estimator for maximizing the chance of achieving consistent estimation and providing the most accurate finite-sample estimation. In this section, we provide a result that updating the initial super-learner estimator through TMLE yields a targeted estimate of the P_0-optimal surrogate that captures all information about the clinical outcome in the following sense. If one would use this targeted estimate as the actual outcome of interest in the current study, and one estimates the average treatment effect on this surrogate with an efficient TMLE based on the reduced data in the current study that ignores the clinical outcome, then this TMLE estimate is an efficient estimator of the average treatment effect on the actual clinical outcome,

1.5.1 The targeted estimate of the P_0-optimal surrogate using TMLE

Suppose Y is binary or continuous in $(0, 1)$. Let ψ_n be the super-learner estimator of $\psi_0(W, A, S) = E_0(Y \mid W, A, S)$. Consider the submodel $\text{Logit}\psi_n^{\#}(\epsilon) = \text{Logit}\psi_n^{\#} + \epsilon H_{g_n}$, where $H_{g_n}(W, A, S) = (2A - 1)/g_n(A \mid W)$, and g_n is an estimator of $g_0(A \mid W)$. In a randomized clinical trial (RCT), we might set $g_n = g_0$. Let $\epsilon_n = \arg\min_\epsilon P_n L(\psi_n^{\#}(\epsilon))$ be the MLE, where P_n is the empirical distribution of the n observations and

$$L(\psi)(O) = -\left\{\psi(W, A, S)^Y (1 - \psi(W, A, S))^{1-Y}\right\} \tag{1.4}$$

is the log-likelihood loss function. This ϵ_n is easily calculated with a standard univariate logistic regression incorporating an offset. Let $\psi_n^{\#} = \psi_n^{\#}(\epsilon_n)$ be the corresponding estimator

of ψ_0, which is a TMLE (indicated by the superscript #) for reasons that we summarize below. This estimator $\psi_n^{\#}$ does not have a closed-form solution unless the super-learner library is very simple, but this does not matter for the purpose of achieving a most useful surrogate given its values are easily calculated.

TMLE is a general approach that allows one to target an initial estimator of a data distribution or parameter thereof in such a way that this targeted version will solve a user-supplied estimating equation [65]. In a typical application of TMLE one targets the initial estimator to solve the efficient influence curve equation for the target parameter of interest so that the resulting substitution estimator is an asymptotically efficient estimator. In the above case, we depart from this objective, instead using the TMLE solely as a technical procedure to make the estimator solve the equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} H_{g_n}(W_i, A_i)(Y_i - \psi_n^{\#}(W_i, A_i, S_i)), \tag{1.5}$$

which is the crucial equation that we will need later for a main result (Theorem 4) that a TMLE of the average treatment effect (ATE) on $\theta_{\psi_n^{\#}}$ is also a TMLE of the ATE on Y and is thus asymptotically linear and efficient for the ATE on Y.

1.5.2 The targeted estimate of the P_0-optimal surrogate is optimal in the current study.

Suppose we use this $\psi_n^{\#}(W, A, S)$ in place of the final outcome Y, and, based on the reduced data $(W_i, A_i, \psi_n^{\#}(W_i, A_i, S_i))$, $i = 1, \ldots, n$ in our current study, compute the TMLE $\theta_{\psi_n^{\#}}^{TMLE}$ of the ATE $\theta_{\psi_n^{\#}} = E_0(\psi_n^{\#}(W, 1, S_1)) - E_0(\psi_n^{\#}(W, 0, S_0))$ on $\psi_n^{\#}(W, A, S)$. This TMLE is an efficient estimator of this data adaptive target parameter $\theta_{\psi_n^{\#}}$, but we are really interested in estimating the ATE $\theta_0 = E_0(Y_1 - Y_0)$ on the clinical outcome Y. So how much information did we lose by replacing the outcome Y by this estimated surrogate outcome $\psi_n^{\#}(W, A, S)$ for the sake of estimation of the desired parameter θ_0?

To answer this question, we first define both the reduced data TMLE $\theta^{TMLE}_{\psi^{\#}_n}$ of $\theta_{\psi^{\#}_n}$ and the TMLE $\widetilde{\theta}^{TMLE}_n$ of θ_0 based on the full data including Y. From this it will be clear that $\theta^{TMLE}_{\psi^{\#}_n}$ is an actual TMLE of θ_0 based on $O = (W, A, S, Y)$ so that its asymptotic properties follow from the well-known theory for TMLE.

TMLE $\tilde{\theta}^{TMLE}_n$ of $E_0(Y_1 - Y_0)$ based on $O = (W, A, Y)$ (without using S): First, we note that an efficient estimator of $EY_1 - EY_0$ can ignore S so that it suffices to work with (W, A, Y). Let $\bar{Q}^0_n$ be an initial estimator of $\bar{Q}_0 = E_0(Y \mid W, A)$ based on (W, A, Y). Let $L(\bar{Q})$ be the log-likelihood loss (1.4), $\text{Logit}\bar{Q}^0_n(\epsilon) = \text{Logit}\bar{Q}^0_n + \epsilon H_{g_n}$ be the least favorable submodel, and $\tilde{\epsilon}_n = \arg\min_\epsilon P_n L(\bar{Q}^0_n(\epsilon))$ be the MLE of the fluctuation parameter ϵ. The TMLE of $\bar{Q}_0$ is defined as $\bar{Q}^{\#1}_n = \bar{Q}^0_n(\tilde{\epsilon}_n)$ and the TMLE of the average treatment effect $E_0(Y_1 - Y_0)$ is given by $\tilde{\theta}^{TMLE}_n = \frac{1}{n}\sum_{i=1}^n \{\bar{Q}^{\#1}_n(W_i, 1) - \bar{Q}^{\#1}_n(W_i, 0)\}$. Due to the TMLE-update step we have that $\bar{Q}^{\#1}_n$ solves the score equation

$$0 = \frac{1}{n}\sum_{i=1}^n H_{g_n}(W_i, A_i)(Y_i - \bar{Q}^1_n(W_i, A_i)), \tag{1.6}$$

and, as a result, the TMLE solves the efficient influence curve equation:

$$0 = \frac{1}{n}\sum_{i=1}^n \widetilde{D}^{eff}(\bar{Q}^{1\#}_n, g_n)(W_i, A_i, Y_i) = 0 \tag{1.7}$$

with $\widetilde{D}^{eff}(\bar{Q}^{1\#}_n, g_n)(W_i, A_i, Y_i) = \widetilde{D}^{eff,1}(\bar{Q}^{1\#}_n, g_n)(W_i, A_i, Y_i) - \widetilde{D}^{eff,0}(\bar{Q}^{1\#}_n, g_n)(W_i, A_i, Y_i)$, where $\widetilde{D}^{eff,a}(\bar{Q}^{1\#}_n, g_n)(W_i, A_i, Y_i) = (I(A_i = a)/g_n(a|W_i))(Y_i - \bar{Q}^{1\#}_n(W_i, a)) + \bar{Q}^{\#1}_n(W_i, a) - \tilde{\theta}^{TMLE,a}_n$ provides the TMLE $\tilde{\theta}^{TMLE,a}_n$ for EY_a as the solution to $0 = \frac{1}{n}\sum_{i=1}^n \widetilde{D}^{eff,a}(\bar{Q}^{1\#}_n, g_n)(W_i, A_i, Y_i) = 0$. We provide the treatment-specific TMLEs because in the application we estimate non-additive difference treatment effects. These equations are standard TMLE equations (e.g., defined in van der Laan and Rose, 2011, p. 527–529 [65]), and are the basis for the double robustness and asymptotic efficiency of the TMLEs.

TMLE $\theta^{TMLE}_{\psi^{\#}_n}$ of the ATE $\theta_{\psi^{\#}_n}$ on the surrogate outcome $\psi^{\#}_n$ based on $O^r = (W, A, \theta_{\psi^{\#}_n}(W, A, S))$: This TMLE is the same as the TMLE above but with Y replaced by $\psi^{\#}_n(W, A, S)$. Thus, one first regresses $\psi^{\#}_n(W_i, A_i, S_i)$ on (W_i, A_i) to obtain an estimator of $\bar{Q}_0(W, A) = E_0(\psi_0(W, A, S) \mid W, A) = E_0(Y \mid W, A)$, where one might again use super-learning. Let us denote this estimator with $\bar{Q}^{\#}_n$. This is nothing else than an estimator of $\bar{Q}_0(W, A) = E_0(E_0(Y \mid W, A, S) \mid W, A)$, which estimates the inner expectation $E_0(Y \mid W, A, S)$ with $\psi^{\#}_n$ and then estimates the outer expectation with a regression of $\psi^{\#}_n$ on (W, A). One now defines the submodel $\text{Logit}\bar{Q}^{\#}_n(\epsilon) = \text{Logit}\bar{Q}^{\#}_n + \epsilon H_{g_n}$, and defines $\epsilon_{n1} = \arg\min_\epsilon \sum_{i=1}^n L_1(\bar{Q}^{\#}_n(\epsilon))(O^r_i)$, where $L_1(\bar{Q})(O^r) = -\Big\{\bar{Q}(W, A)^{\psi^{\#}_n(W,A,S)}$ $(1 - \bar{Q}(W, A))^{1-\psi^{\#}_n(W,A,S)}\Big\}$. This TMLE $\bar{Q}^{\#}_n = \bar{Q}_n(\epsilon_{n1})$ solves the following score equation (analog to (1.6)):

$$0 = \frac{1}{n}\sum_{i=1}^n H_{g_n}(W_i, A_i)(\psi^{\#}_n(W_i, A_i, S_i) - \bar{Q}_n(\epsilon_{n1})). \tag{1.8}$$

The TMLE $\theta^{TMLE}_{\psi^{\#}_n}$ of $\theta_{\psi^{\#}_n}$ is now the substitution estimator

$$\theta^{TMLE}_{\psi^{\#}_n} = \frac{1}{n}\sum_{i=1}^n \{\bar{Q}_n(\epsilon_{n1})(W_i, 1) - \bar{Q}_n(\epsilon_{n1})(W_i, 0)\}.$$

Now we utilize the fact that $\psi^{\#}_n$ was targeted so that it solves the equation (1.5). Equation (1.5) combined with the score equation (1.8) implies that $\bar{Q}^{\#}_n$ solves

$$0 = \frac{1}{n}\sum_{i=1}^n H_{g_n}(W_i, A_i)(Y_i - \bar{Q}_n(\epsilon_{n1})). \tag{1.9}$$

Thus, this TMLE $\theta^{TMLE}_{\psi^{\#}_n}$, defined as the solution to

$$0 = \frac{1}{n}\sum_{i=1}^n D^{eff}(\bar{Q}^{\#}_n, g_n)(W_i, A_i, Y_i) = 0 \tag{1.10}$$

with $D^{eff}(\bar{Q}_n^{\#}, g_n)(W_i, A_i, Y_i) = D^{eff,1}(\bar{Q}_n^{\#}, g_n)(W_i, A_i, Y_i) - D^{eff,0}(\bar{Q}_n^{\#}, g_n)(W_i, A_i, Y_i)$, also solves the efficient influence curve equation (1.7), where $D^{eff,a}$ is the same as $\widetilde{D}^{eff,a}$ except $\bar{Q}_n^{\#1}$ is replaced with $\bar{Q}_n^{\#}$ and $\widetilde{\theta}_n^{TMLE,a}$ with $\theta_{\psi_n^{\#}}^{TMLE,a}$. (And the separate TMLEs $\theta_{\psi_n^{\#}}^{TMLE,a}$ for $E_0(\psi_n^{\#}(W, A = a, S_a))$ for $a = 0, 1$ are obtained in exactly parallel fashion.) As a consequence, $\theta_{\psi_n^{\#}}^{TMLE}$ is an asymptotically efficient estimator of $E_0(Y_1 - Y_0)$ under the usual conditions for a TMLE. When this score equation (1.9) for ϵ is unique, $\bar{Q}_n(\epsilon_{n1}) = \bar{Q}_n(\epsilon_{n2})$ showing that $\theta_{\psi_n^{\#}}^{TMLE}$ is an actual TMLE of $E_0(Y_1 - Y_0)$ based on the original data (W, A, S, Y), with the only twist that it uses a special initial estimator $\bar{Q}_n$ of $\bar{Q}_0$ (as discussed above). This proves that $\theta_{\psi_n^{\#}}^{TMLE}$– which we defined as a TMLE of the treatment effect on the estimated optimal surrogate – is also a double robust efficient substitution estimator of the clinical treatment effect of interest $E_0(Y_1 - Y_0)$ based on the full data $O = (W, A, S, Y)$ in model $\mathcal{M}$. We collate this as Theorem 4.

THEOREM 4: *Consider the estimator* $\psi_n^{\#}$ *of* $\psi_0 = E_0(Y \mid W, A, S)$ *and the TMLE* $\theta_{\psi_n^{\#}}^{TMLE}$ *of* $\theta_{\psi_n^{\#}} = E_0(\psi_n^{\#}(W, 1, S_1) - \psi_n^{\#}(W, 0, S_0))$ *based on* $(W_i, A_i, \psi_n^{\#}(W_i, A_i, S_i))$, $i = 1, \ldots, n$. *Let* $\theta_0 = E_0(Y_1 - Y_0)$. *Let* $Q_0 = (\bar{Q}_0, Q_{W,0})$, *where* $Q_{W,0}$ *is the probability distribution of* W *under* P_0. *Let* $\tilde{D}^{eff}(Q_0, g_0)(O) = H(g_0)(W, A)(Y - \bar{Q}_0(W, A)) + \bar{Q}_0(W, 1) - \bar{Q}_0(W, 0) - \theta(Q_0)$ *be the efficient influence curve of* $E_0(Y_1 - Y_0)$ *based on the data* $O = (W, A, S, Y) \sim P_0 \in \mathcal{M}$. *Let* $Q_n^{\#} = (\bar{Q}_n^{\#}, Q_{W,n})$ *and let* $\| f \|_{P_0} = \sqrt{\int f(o)^2 dP_0(o)}$. *Assume 1)* $\tilde{D}^{eff}(Q_n^{\#}, g_n)$ *falls in a* P_0*-Donsker class with probability tending to 1; 2)* $\| \bar{Q}_n^{\#} - \bar{Q}_0 \|_{P_0} \| g_n - g_0 \|_{P_0} = o_P(1/\sqrt{n})$ *(so in an RCT, this only requires* $\| \bar{Q}_n^{\#} - \bar{Q}_0 \|_{P_0} \to 0$ *in probability); 3) for some* $\delta > 0$ $\min_{a \in \{0,1\}} g_0(a \mid W) > \delta > 0$ *with probability 1. Then* $\theta_{\psi_n^{\#}}^{TMLE} - \theta_0 = (P_n - P_0)\tilde{D}^{eff}(Q_0, g_0) + o_P(1/\sqrt{n})$. *Thus,* $\theta_{\psi_n^{\#}}^{TMLE}$ *is an efficient estimator of* θ_0 *based on* $O = (W, A, S, Y)$ *in model* $\mathcal{M}$.

Moreover, even though $\theta_{\psi_n^{\#}}^{TMLE}$ is based on a reduced data structure, it is asymptotically linear with influence curve equal to that of the TMLE $\tilde{\theta}_n^{TMLE}$ of $E_0(Y_1 - Y_0)$ based on the observed data (W, A, Y). This is an important result since it establishes that in our original

study the estimated optimal surrogate carries as much information as the outcome itself for the sake of estimation of the average clinical treatment effect. This means that a Wald $(1-\alpha)\%$ confidence interval for $\theta_{\psi_n^\#}$ based on $\theta_{\psi_n^\#}^{TMLE}$ is also a $(1-\alpha)\%$ confidence interval for $\theta_0 = E_0(Y_1 - Y_0)$ and is as narrow as a $(1-\alpha)\%$ confidence interval based on an efficient estimator of θ_0 using (W, A, S, Y).

This result may be surprising given that the estimated optimal surrogate is based on the reduced data. In fact, if a super-learner estimator were used as the estimated optimal surrogate, without targeting the estimator, then the TMLE $\theta_{\psi_n^\#}^{TMLE}$ would not be efficient for $E_0(Y_1 - Y_0)$. Specifically, the bias of a super-learner fit is larger than the inverse of root-n and this bias translates into the same order of bias for the ATE on Y. The key to achieving efficiency is therefore to use a targeted super-learner fit of the optimal surrogate designed so that the TMLE of the ATE on this targeted estimate is in fact an asymptotically linear estimator of the ATE on Y. However, this targeting is only possible if we use the actual observed outcomes Y, and the targeting is specific for the current data generating experiment and thus the TMLE of the average treatment effect on our targeted surrogate based on a new future study would not result in an asymptotically efficient estimator of the ATE on Y. Nevertheless, it is an appealing property to have a surrogate such that if we use it in a certain way in our current study it yields an asymptotically efficient estimator of the average clinical treatment effect in the current study.

1.6 Application to Two Simulated Dengue Vaccine Efficacy Trials

Two randomized, double-blinded, placebo-controlled, multicenter, Phase 3 trials of the identical recombinant, live, attenuated, tetravalent dengue vaccine (CYD-TDV) versus placebo were conducted in Asia [9] and Latin America [70], respectively. These trials– named CYD14 and CYD15– randomized 10,275 and 20,869 children, respectively, in 2:1 allocation to vaccine:placebo, with immunizations administered at months 0, 6, and 12. The primary analyses

assessed vaccine efficacy (VE) against symptomatic, virologically confirmed dengue (VCD) occuring at least 28 days after the third immunization through to the Month 25 visit. Based on a proportional hazards model, estimated VE was 56.5% (95% CI 43.8–66.4) for CYD14 and 64.7% (95% CI 58.7–69.8) for CYD15.

The trials measured neutralizing antibody titers to each of the four dengue serotypes contained in the CYD-TDV vaccine, at baseline and at Month 13. An objective is to develop a surrogate endpoint for VCD based on these titer variables. These titers were measured via case-cohort sampling, and we illustrate the new framework with complete pseudo CYD14 and CYD15 data sets generated by sampling and simulations from the real data sets (see Section 1.10); we refer to these simulated data sets as simCYD14 and simCYD15. SimCYD14 and simCYD15 each consist of baseline covariates W (age, sex, four baseline titers), treatment A (1=vaccine, 0=placebo), S (the four Month 13 titers), and Y the indicator of occurrence of the VCD endpoint between Month 13 and Month 25 (Section 1.11 describes the data sets in further detail). We treat simCYD14 as the current trial and simCYD15 as the future trial. In Chapter 2 we develop optimal surrogate methodology for case-cohort sampling designs and apply the methodology to the CYD14 and CYD15 data from the original clinical trials.

We first obtain the targeted estimated optimal surrogate $\psi_n^{TMLE}(W, A, S)$ for the simCYD14 trial, thus obtaining TMLEs $\theta_{\psi_n}^{TMLE,a}$ of each $E_0(\psi_n^{TMLE}(W, a, S_a))$ and $\theta_n^{TMLE} = 1 - \theta_{\psi_n}^{TMLE,1}/\theta_{\psi_n}^{TMLE,0}$ of $\theta_{\psi_n} = 1 - E_0(\psi_n^{TMLE}(W, 1, S_1))/E_0(\psi_n^{TMLE}(W, 0, S_0))$. Second, we calculate the $\psi_n^{TMLE}(W^*, A^*, S^*)$ surrogate outcome values for the n^* simCYD15 participants (with $\psi_n^{TMLE}(\cdot)$ from simCYD14), and, based on the simCYD15 data $(W_i^*, A_i^*, S_i^*, \psi_n(W_i^*, A_i^*, S_i^*))$, $i = 1, \cdots, n^*$, estimate the treatment-specific surrogate means $\theta_{\psi_n}^a(P) = E_P\left[E_P(\psi_n(W^*, a, S^*) \mid A^* = a, W^*)\right]$ for $a \in \{0, 1\}$ and the vaccine efficacy on the surrogate $\theta_{\psi_n}(P) = VE_{\psi_n}(P) = 1 - \theta_{\psi_n,1}^1(P)/\theta_{\psi_n}^0(P)$ [estimated by $\theta_{\psi_n}^{TMLE,a}(P) = (1/n^*)\sum_{i=1}^{n^*} \psi_n^{TMLE}(W_i^*, a, S_i^*)$ and $\theta_{\psi_n}^{TMLE}(P) = VE_{\psi_n}^{TMLE}(P) = 1 - \theta_{\psi_n}^{TMLE,1}(P)/\theta_{\psi_n}^{TMLE,0}(P)$]. Lastly, because we know the Y^* values in simCYD15, we compare the targeted surrogate based estimates $\theta_{\psi_n}^{TMLE,1}(P)$,

$\theta_{\psi_n}^{TMLE,0}(P)$, and $VE_{\psi_n}^{TMLE}(P)$ to the TMLEs of $E_P(Y_0^*)$, $E_P(Y_1^*)$, and $\theta_P^* = VE_P^* = 1 - E_P(Y_1^*)/E_P(Y_0^*)$, respectively, based on the simCYD15 data (A_i^*, Y_i^*), to check how well the estimated optimal surrogate can be used to estimate the clinical risk and treatment effect parameters in the new setting. Wald 95% confidence intervals for the $\theta_{\psi_n}^a(P)$ and $E_P(Y^*)$ parameters are calculated based on the influence functions, and then for $VE_{\psi_n}(P)$ and VE_P^* using the delta method.

1.6.1 Targeted super-learner estimate of $\psi_0 = E_0(Y|W, A, S)$ in the simCYD14 trial

We applied super-learner with 10-fold cross-validation, separately for the vaccine and placebo groups. Table 1.1 shows the input variables and 48 learners. Figure 1.1 shows point and 95% CI estimates of the cross-validated MSEs [62] for each individual learning algorithm as well as for discrete super-learner and super-learner. The model that uses a generalized additive model with a 4 degrees of freedom (df) smoothing spline for each of the 8 neutralization titer variables and includes age and gender (labeled gam4) performs best (with the lowest CV-MSE) for each treatment arm, and thus is selected by the discrete super-learner, and the super-learner has almost identical CV-MSE. Classification accuracy is better for the vaccine than placebo group, e.g., with cross-validated MSE of the super-learner 0.014 (95% CI 0.012–0.017) and 0.032 (95% CI 0.027–0.037), respectively. Table 1.2 shows the gam4 model that was selected by the discrete super-learner, including an expression for the model.

Next, the TMLE $\psi_n^{TMLE}(W, A, S)$ was obtained as discussed in Section 1.5.1. To study how well this estimated optimal surrogate classifies VCD in simCYD14, Figure 1.2(a) shows empirical reverse cdf plots of $\psi_n^{TMLE}(W_i, A_i = a, S_i)$ by treatment group $a \in \{0, 1\}$ and VCD case-control outcome $y \in \{0, 1\}$ for simCYD14 data, again showing better classification in the vaccine group. Based on $\psi_n^{TMLE}(W, A, S)$, $\widehat{E}_0(Y_1) = \theta_{\psi_n}^{TMLE,1} = 0.018$ (95% CI 0.014–0.023), $\widehat{E}_0(Y_0) = \theta_{\psi_n}^{TMLE,0} = 0.037$ (95% CI 0.023–0.060), and $\widehat{VE}_0 = \theta_n^{TMLE} = 52\%$ (95% CI 41–66)– this inference on VE_0 is close to the original primary study result of $\widehat{VE}_0 = \tilde{\theta}_n^{TMLE}$

= 56% (95% CI 44–66). potentially supporting the validity of the proposed conditions as discussed in Section 1.3.4.

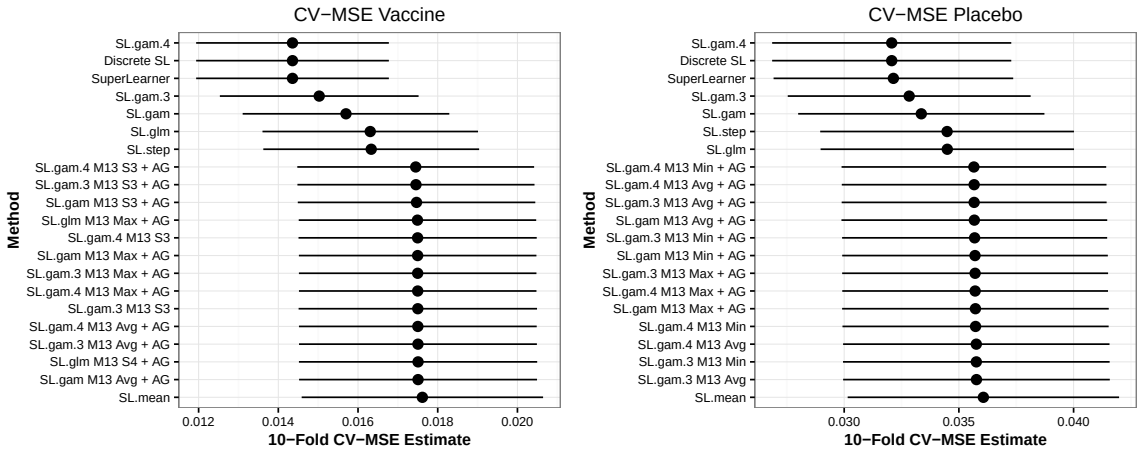

Figure 1.1: Point and 95% confidence interval estimates of cross-validated mean squared error (CV-MSE) for the vaccine and placebo groups of the simCYD14 trial, for individual learners, discrete super-learner, and super-learner.

1.6.2 Applying $\psi_n^{TMLE}(\cdot)$ developed from simCYD14 data to the simCYD15 trial

Table 1.3 compares estimates of the surrogate and clinical parameters in simCYD15. The results show a higher estimate of $VE_{\psi_n}(P)$ based on the surrogate than of VE_P^* based directly on the clinical endpoint, with $VE_{\psi_n}(P) = 68\%$ (95% CI 58–81) and $\widehat{VE}_P^* = \tilde{\theta}_{n*}^{TMLE} =$

Table 1.1: Input variables and learners (62 algorithms/learners) used in the super-learner for the simCYD14 dengue vaccine efficacy trial.

Input Variables	
W	Baseline demographics age (range 2–14 years), sex
W	Baseline titers to the 4 serotypes inside the CYD-TDV vaccine, average, minimum, and maximum of the 4 titers
S	Month 13 titers to the 4 serotypes inside the CYD-TDV vaccine, average, minimum, and maximum of the 4 titers
Learners	**Boldfaced Courier-font learning algorithms (e.g., `SL.mean`) are built into the SuperLearner R package available at CRAN**
`SL.mean`	$E_0(Y\|W, A = a, S)^a = \beta_a$ for $a \in \{0, 1\}$
`SL.glm`	Logistic regression with all input variables
`SL.step`[b]	Best logistic regression model by AIC through a step-wise search
`SL.gam`	gam[c] for W & S inputs; all titer variables each with 2 df
`SL.gam.3`	gam for W & S inputs; all titer variables each with 3 df
`SL.gam.4`	gam for W & S inputs; all titer variables each with 4 df
M13 Sk[d]	`SL.glm`, `SL.gam`, `SL.gam.3`, `SL.gam.4` with only Month 13 serotype k titers
M13 Avg	`SL.glm`, `SL.gam`, `SL.gam.3`, `SL.gam.4` with only Month 13 average titers
M13 Min, Max	`SL.glm`, `SL.gam`, `SL.gam.3`, `SL.gam.4` with only Month 13 Min or Max titers
M13 Sk + AG	`SL.glm`, `SL.gam`, `SL.gam.3`, `SL.gam.4` with Month 13 serotype k titers + (age, gender)
M13 Avg + AG	`SL.glm`, `SL.gam`, `SL.gam.3`, `SL.gam.4` with Month 13 average titers + (age, gender)
M13 Min, Max + AG	`SL.glm`, `SL.gam`, `SL.gam.3`, `SL.gam.4` with Month 13 Min or Max titers + (age, gender)
`Discrete SL`	van der Laan, Polley, and Hubbard (2007)
`Super Learner (SL)`	van der Laan, Polley, and Hubbard (2007)

[a] All learners were fit separately for each treatment group $A = a$ for $a \in \{0, 1\}$ as described in Section 1.6.1. This is explicitly stated here for `SL.mean`, and can be assumed for all other learners.

[b] The step-wise search begins with the full (all input variables) model. On the first step, it fits all remove-one-variable models and then removes one variable based on lowest AIC of the remove-one models. For next and subsequent steps, all models dropping or adding one variable are checked and a variable is dropped or added based on the model with lowest AIC. This procedure is repeated until no add-one-variable or drop-one-variable model has lower AIC than the current model.

[c] Generalized additive model of Hastie and Tibshirani (1990).

[d] k refers to the respective serotype (1, 2, 3, or 4).

Table 1.2: Discrete super-learner models (gam4) for the vaccine and placebo groups of the simCYD14 trial. For each treatment group, the fitted model is $\text{logit}(\widehat{P}(Y = 1)|w, x) = \widehat{\beta}_0 + \widehat{\beta}_1^T w + \widehat{\beta}_2^T s$ where w is age, gender, and the six baseline serotype variables and s is the four Month 13 serotype variables, and $s(\cdot, 4)$ denotes a smoothing spline with 4 degrees of freedom, with the estimated splines plotted in Supplementary Figure 5.

Model Term[a]	Coefficient	Odds Ratio	2-Sided P-value
Vaccine Model			
s(M0_SeroType1, 4)	1.33	3.77	8.2e-08
s(M0_SeroType2, 4)	0.54	1.72	0.84
s(M0_SeroType3, 4)	0.43	1.53	0.51
s(M0_SeroType4, 4)	1.31	3.71	3.5e-11
s(M0_MinSeroTiter, 4)	-0.84	0.43	0.9
s(M0_MaxSeroTiter, 4)	-0.71	0.49	0.0016
s(M13_SeroType1, 4)	-0.33	0.72	3.3e-06
s(M13_SeroType2, 4)	-0.51	0.60	0.0023
s(M13_SeroType3, 4)	-0.41	0.66	0.017
s(M13_SeroType4, 4)	-0.43	0.65	0.021
Age 6–11[b]	0.15	1.16	0.25
Age 12–14[b]	-0.63	0.53	0.049
Male	-0.66	0.52	0.00041
Placebo Model			
s(M0_SeroType1, 4)	0.81	2.26	0.7
s(M0_SeroType2, 4)	0.09	1.09	1.4e-09
s(M0_SeroType3, 4)	-1.56	0.21	6.1e-13
s(M0_SeroType4, 4)	0.72	2.05	0.17
s(M0_MinSeroTiter, 4)	-1.44	0.24	1.3e-05
s(M0_MaxSeroTiter, 4)	-0.36	0.69	0.87
s(M13_SeroType1, 4)	0.30	1.35	2.6e-05
s(M13_SeroType2, 4)	-0.07	0.93	0.3
s(M13_SeroType3, 4)	1.01	2.75	3.7e-06
s(M13_SeroType4, 4)	-0.46	0.63	0.021
Age 6–11[b]	0.50	1.64	0.00095
Age 12–14[b]	-0.17	0.84	0.46
Male	-0.25	0.78	0.14

[a] s(. . .,k) indicates a smoothing spline fit with k degrees of freedom.
[b] The reference age category is 2–5 year olds.

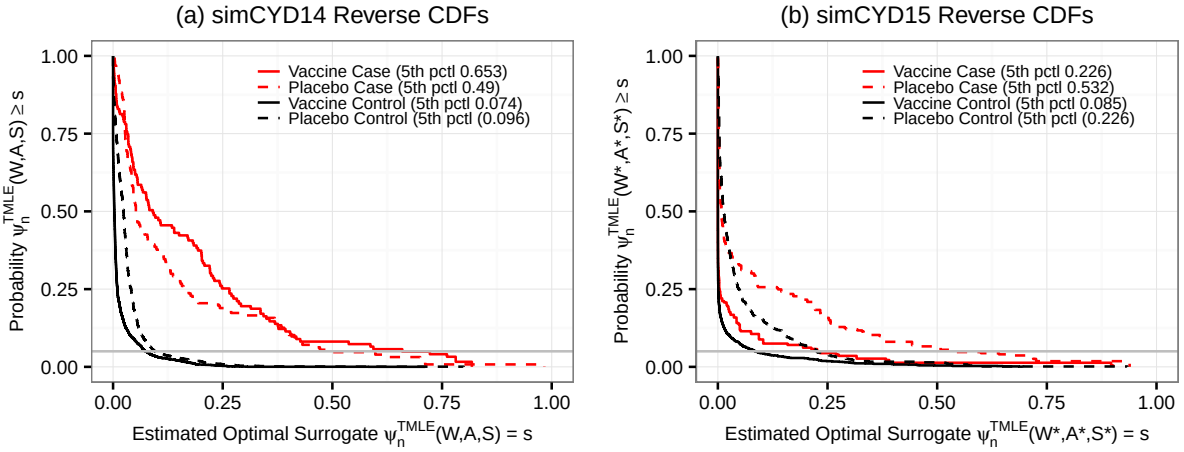

Figure 1.2: (a) Empirical reverse cumulative distribution functions (cdfs) of the estimated optimal surrogate $\psi_n^{TMLE}(W_i, A_i = a, S_i)$ for the simCYD14 trial by vaccine/placebo assignment $A = a \in \{0, 1\}$ and dengue outcome case/control status $Y = y \in \{0, 1\}$. (b) Empirical reverse cdfs of $\psi_n^{TMLE}(W_i^*, A_i^* = a, S_i^*)$ for simCYD15 participants by vaccine/placebo assignment $A^* = a \in \{0, 1\}$ and dengue outcome case/control status $Y^* = y \in \{0, 1\}$, where $\psi_n^{TMLE}(\cdot)$ was estimated from the simCYD14 trial data. The results show that the surrogate better classifies dengue outcomes of participants in the original trial than in the new trial, as expected.

59% (95% CI 54–65). The discrepancy is probably due to violations of the assumptions of Theorems 1 and 2– as detailed in Section 1.11.1 diagnostics suggest some violations of Equal Conditional Means and the contained-support assumption does not perfectly hold because simCYD14 included 2–14 year olds whereas simCYD15 included 9–16 year olds. One simple way to lessen the bias would be to repeat the analysis restricting the simCYD15 study to 9–14 year olds, thereby making the contained-support assumption hold. Figure 1.2(b) shows empirical reverse cdf plots of $\psi_n^{TMLE}(W_i^*, A_i^* = a, S_i^*)$ for each treatment $a \in \{0,1\}$ by case-control status $y \in \{0,1\}$ in simCYD15, showing diminution of classification accuracy of the surrogate built on simCYD14 for the new study simCYD15 (as expected).

Table 1.3: Comparison of inferences on the surrogate parameters in which $\theta^a_{\psi_n}(P) \equiv E_P[E_P(\psi_n(W^*, a, S^*) \mid A^* = a, W^*)]$ for each $a \in \{0,1\}$ and $\theta_{\psi_n}(P) = VE_{\psi_n}(P) = 1 - \theta^1_{\psi_n}(P)/\theta^0_{\psi_n}(P)$ based on $(W^*, A^*, S^*, \psi_n(W^*, A^*, S^*))$ versus inferences on the clinical dengue endpoint parameters $E_P(Y_a^*)$ and $\theta_P^* = VE_P^* = 1 - E_P(Y_1^*)/E_P(Y_0^*)$ in simCYD15.

Surrogate Parameters Based on the TMLE of the Optimal Surrogate TMLE θ_n^{TMLE}		Clinical Parameters Based on the TMLE $\tilde{\theta}_{n^*}^{TMLE}$	
$\theta^1_{\psi_n}(P)$	0.017 (95% CI 0.014–0.020)	$E_P(Y_1^*)$	0.017 (95% CI 0.014–0.019)
$\theta^0_{\psi_n}(P)$	0.053 (95% CI 0.040–0.069)	$E_P(Y_0^*)$	0.040 (95% CI 0.036–0.045)
$VE_{\psi_n}(P)$	68% (95% CI 58–81)	VE_P^*	61% (95% CI 54–67)

1.6.3 Bridging dengue vaccine efficacy from 9–16 year olds to 18–45 year olds.

The CYD-TDV vaccine has been licensed for the age range included in the CYD14 and CYD15 Phase 3 trials: 9–16 year olds. It is of interest to also license the vaccine for adults. However, for ethical reasons the previous Phase 3 results preclude conducting a placebo-controlled efficacy trial in adults. Accordingly, a randomized Phase 1 trial of the CYD-TDV vaccine ($A^* = 1$) versus placebo ($A^* = 0$) was conducted in 18–45 year olds in a

CYD14 study country (Vietnam), and the same variables (W^*, A^*, S^*) were measured. The estimated optimal surrogate built from CYD14 9–16 year olds could be used to estimate each $E_P(Y^*|A^* = a)$ for $a \in \{0, 1\}$ in 18–45 year olds. Validity of these estimates would depend on the assumptions of Theorems 2 and 3, which would require careful scrutiny.

1.7 Simulation Studies

We conduct a small simulation study to illustrate that the targeted estimated optimal surrogate will generally provide unbiased estimation of $\theta_0 = E_0(Y_1 - Y_0)$ in the original trial, for any distribution of (W, A, S, Y), whereas in contrast a proportion of treatment effect explained based approach that has been popular in practice does not. We then evaluate how well the surrogate built from the original study can be used to estimate θ_0 in a new study that only measures (W, A, S), when the Equal Conditional Means assumption fails.

1.7.1 Data generating distribution

Building upon an example in VanderWeele (2013), we simulate data illustrating the surrogate paradox. The data set is comprised of an outcome Y, a randomized treatment $A \in \{0, 1\}$, and 10 candidate surrogates S^k, each with three levels $S^k \in \{0, 1, 2\}$ for $k = 1, \ldots, 10$. For each k the joint potential outcomes S^k_a for $a \in \{0, 1\}$ have the following distribution: $P(S^k_1 = 0, S^k_0 = 0) = P(S^k_1 = 1, S^k_0 = 1) = P(S^k_1 = 2, S^k_0 = 2) = 0.1$, $P(S^k_1 = 1, S^k_0 = 0) = 0.5$, and $P(S^k_1 = 1, S^k_0 = 2) = 0.2$. The outcome $Y = \sum_{k=1}^{3} \left[0.1 * k * I(S^k = 1) + I(S^k = 2)\right] + \epsilon_Y$, where $\epsilon_Y \sim N(0, 0.1^2)$. In this setting $\theta_0 = E_0(Y_1 - Y_0) = -0.18$, whereas $E_0(S^k_1 - S^k_0) = 0.3$ for each k, such that the surrogate paradox occurs for each k.

Methods for estimating θ_0 based on a surrogate

The estimated optimal surrogate $\psi^{TMLE}_n(A, S)$ and the resulting estimate θ^{TMLE}_n of θ_0 are calculated as for the first example. We compare performance of θ^{TMLE}_n to an alternative

approach that estimates the Proportion of the Treatment Effect Captured (PCS), as proposed by Kobayashi and Kuroki [35], by each of the ten candidate surrogate endpoints to select the best single surrogate variable as the one that maximizes the estimated PCS, which we refer to as S^{PCSopt}. Specifically, for each of 100 bootstrapped data sets, the index k maximizing the estimated PCS in a linear regression model of Y on $I(S^k = 1)$ and $I(S^k = 2)$ was selected, and S^{PCSopt} was taken to be the S^k most frequently selected. Then θ_0 was estimated by θ_n^{PCSopt} defined as the difference in average predicted Y values for group $a = 1$ minus $a = 0$ in the fitted model $\widehat{E}_0(Y|S_{opt}^{\text{PCS}} = s, A = a) = \widehat{\beta}_0 + \widehat{\beta}_1 * I(s = 1) + \widehat{\beta}_2 * I(s = 2)$. Since a perfect surrogate captures all of the effect of the treatment A (indicated by PCS=1 in the proportion-of-treatment-effect explained paradigm), A was not included in the model. The true PCS values are 0.87, 0.2, and 0.002 for the first three S^k's that are predictive of Y.

Simulation 1: Comparison for estimating θ_0 in the original trial

For each of 200 hundred simulated data sets each with 2000 subjects, θ_0 was estimated based on the SL-TMLE surrogate and the PCS-selected surrogate as described above. Figure 1.3(a) shows the concordance of the surrogate-based estimates of θ_0 and the gold-standard estimates based on the known clinical outcomes, $\tilde{\theta}_n^{TMLE} = \widehat{E}_0(Y_1) - \widehat{E}_0(Y_0)$, where the $\widehat{E}_0(Y_a)$'s are sample averages. The estimates θ_n^{TMLE} much more closely align with the direct Y-based estimates than those based on θ_n^{PCSopt}, with average θ_n^{TMLE} of -0.18 and average θ_n^{PCSopt} of 0.02, compared to the true value $\theta_0 = -0.18$. The surrogate paradox defined by positive θ_n^{PCSopt} occurred for 191 (96%) of 200 repetitions, whereas it never occurred based on θ_n^{TMLE}. The PCS-based approach fails because it is not able to capture the 3-variable relationship from the data generating distribution, with CV-R^2 of -0.01 between the S^{PCSopt}-estimated $\widehat{Y}$ values and the Y values, compared to CV-R^2 of 0.98 from the estimated optimal surrogate.

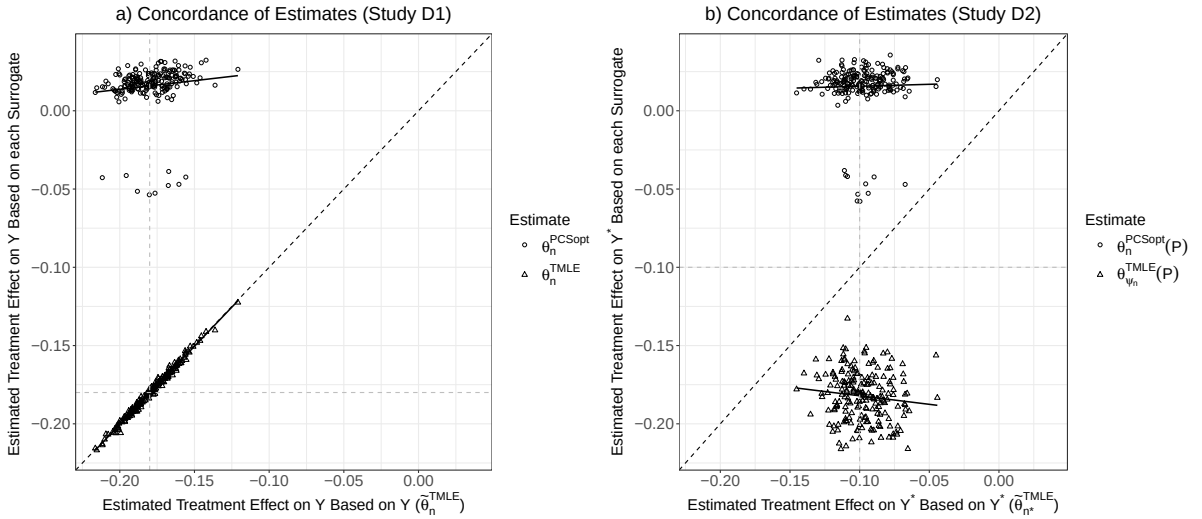

Figure 1.3: (a) For Simulation 1, estimates of $\theta_0 = E_0(Y_1 - Y_0)$ based on two surrogate endpoint approaches [superlearner-TMLE (SL-TMLE) and proportion of treatment effect captured (PCS)] versus estimates based on sample averages of the clinical endpoints Y. For the PCS method, S^{PCSopt} was selected to be S^1 (the best candidate surrogate, with PCS=0.87) in 191 of 200 (95%) data sets. (b) For Simulation 2, estimates of $\theta_P^* = E_P(Y_1^* - Y_0^*)$ for a second trial D2 based on the two surrogate endpoint approaches with surrogates built from the first trial D1, versus estimates based on sample averages of the clinical endpoints Y^*.

1.7.2 Simulation 2: Comparison for estimating θ_P^* in a second trial

Our second simulation generates 200 pairs of data sets (D1, D2) with D1 generated as for Simulation 1 (the original trial) and D2 under a new data generating distribution where Y^* also depends on the fourth candidate surrogate: $Y^* = \sum_{i=1}^4 \left[0.1 * k * I(S^{*k} = 1) + I(S^{*k} = 2)\right] + \epsilon_{Y^*}$, where $\epsilon_{Y^*} \sim N(0, 0.1^2)$. The surrogates $\psi_n^{TMLE}(A, S)$ and $S^{\text{PCSopt}}(A, S)$ are calculated from D1 as in Simulation 1. Then, based on the (A^*, S^*) values in the paired data set D2, surrogate-based estimates of $\theta_{\psi_n}^{TMLE}(P) = \theta_{\psi_n}^1(P) - \theta_{\psi_n}^0(P)$ are calculated as in Section 1.6 and $\theta_n^{\text{PCSopt}}(P) = (1/n_1^*) \sum_{i=1}^{n*} A_i^* \widehat{E}[Y | S_i^{*\text{PCSopt}}, A_i^* = 1] - (1/n_0^*) \sum_{i=1}^{n*} (1 - A_i^*) \widehat{E}[Y | S_i^{*\text{PCSopt}}, A_i^* = 0]$. The D2 data set is chosen such that the Equal Conditional Means assumptions is violated, as depicted in Figure 1.4, which shows that $E_P[Y_a^* | S_a^{*4} = s] - E_{P_0}[Y_a | S_a^4 = s]$ varies widely

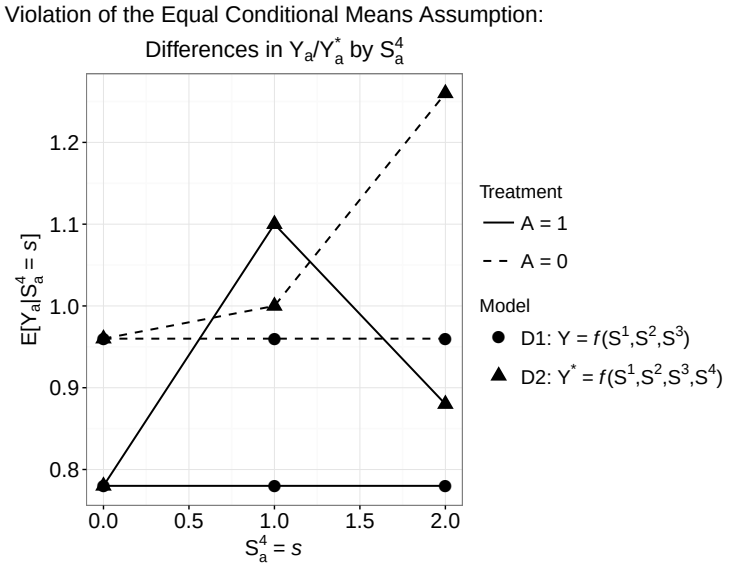

Figure 1.4: Consider two data sets: D1 in which $Y = f(S^1, S^2, S^3) = \sum_{k=1}^{3}[0.1 * k * I(S^k = 1) + I(S^k = 2)] + \epsilon_Y$, and D2 in which $Y^* = f(S^1, S^2, S^3, S^4) = \sum_{k=1}^{4}\left[0.1 * k * I(S^{*k} = 1) + I(S^{*k} = 2)\right] + \epsilon_{Y^*}$ where $\epsilon_Y \sim N(0, 0.1^2)$ and $\epsilon_{Y^*} \sim N(0, 0.1^2)$ (as described in Section 7.4 of the main paper). When comparing the conditional means across values of S^4_a, we see that $E[Y_a|S^4_a = s]$ differs from $E[Y^*_a|S^{*4}_a = s]$ for some values of s (most dramatically for the treatment group $a = 1$ at $s = 1$ and for the control group $a = 0$ at $s = 2$), and thus, the equal conditional means assumption is violated in this example.

in s for each $a \in \{0, 1\}$.

Figure 1.3(b) displays the $\theta_{\psi_n}^{TMLE}(P)$ and $\theta^{\text{PCSopt}}(P)$ estimates versus the gold-standard estimates $\tilde{\theta}_{n*}^{TMLE}$ based on the actual clinical outcomes Y^*. Both approaches demonstrate some bias for estimating θ_P^* (dotted line); however $\theta_{\psi_n}(P)$ does much better at estimating the effect in the correct direction (negative), while $\theta_n^{\text{PCSopt}}(P)$ estimates the effect near zero (the true treatment effect θ_P^* is -0.10, compared to an average estimate $\theta_{\psi_n}(P)$ of -0.18 and an average $\theta_n^{\text{PCSopt}}(P)$ of 0.02). Of the 200 simulation runs, 95% of the PCS-based estimates exhibit the surrogate paradox, compared to 0% for the SL-TMLE method. Therefore, in addition to demonstrating that the Equal Conditional Means assumption is necessary for valid inference on θ_P^* in a new setting, this simulation illustrates that when Equal Conditional Means is majorly violated, the SL-TMLE approach can still preserve some accuracy in bridging the clinical treatment effect to a new setting.

1.8 Discussion of Results

VanderWeele [68] and discussants Joffe [33] and Pearl [45] suggest that a minimal requirement for an intermediate endpoint to be a useful surrogate endpoint is that it avoids the surrogate paradox, which can have disastrous consequences. Yet, VanderWeele [68] shows that commonly used methods for surrogate endpoint evaluation generally do not guarantee avoiding this paradox. The newly proposed approach starts at this minimal requirement, defining the optimal surrogate in a way guaranteed to satisfy the Prentice definition of a valid surrogate and hence avoid the paradox (and hence the estimated optimal surrogate, which can be used as a surrogate endpoint, satisfies the Prentice definition in large samples). As such the new approach responds to Pearl's [45] question: "If we take the negation of the "surrogate paradox" as a criterion for "good" surrogate, why cannot we create a new, formal definition of "surrogacy" that (1) will automatically avoid the paradox?..."

Another new element of the optimal surrogate approach compared to previous approaches

is its use of unbiased supervised statistical learning and targeted minimum loss based estimation to estimate the optimal surrogate. This is fitting to common applications where many baseline covariates and intermediate response endpoints are measured, yet there is large uncertainty about how to best predict the study outcome from these collected data. Moreover, while we have focused on randomized studies, this framework also applies for generating promising candidate surrogates based on observational studies, with all of the results holding under the additional (challenging) assumption that all confounders W of treatment assignment are measured and included in the super-learner.

This chapter considers an ideal setting with no missing data and where the clinical outcome is never observed before the intermediate response endpoints are measured. Future work is of interest to develop estimators of the optimal surrogate that accommodate these issues. Theorems 2 and 3 describe conditions for using the optimal surrogate built from an efficacy trial to confer correct estimation of the clinical treatment effect in a new setting based on this surrogate endpoint without measuring the clinical endpoint, applicable when the new setting entails the same treatment or a new treatment, respectively, and additional research is needed to allow departures from the licensing conditions. Moreover, these theoretical results hold for an infinite original study, such that additional additional research is needed to provide confidence intervals about the clinical treatment effect in a new setting accounting for the error in estimating the optimal surrogate; valid inference is straightforward if the conditional mean is modeled parametrically but not if modeled nonparametrically.

1.9 Connection of the Optimal Surrogate Framework to Other Surrogate Frameworks

Joffe and Greene [34] classified statistical methods for evaluating the validity of candidate surrogate endpoints into four frameworks, which may be referred to as the Prentice replacement endpoint, controlled direct effects, principal stratification, and meta-analysis frame-

works. Prentice [50] catalyzed the field with his definition of a valid surrogate endpoint and operational criteria, as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." Prentice (1989) also provided operational criteria for checking whether an intermediate endpoint satisfies this definition, the most important being the 'full mediation' criterion that the distribution of the clinical endpoint conditional on the surrogate is the same as the distribution of the clinical endpoint conditional on the surrogate and treatment, and many subsequent papers developed methods for evaluating these criteria or related criteria [19, 38, 7, 73, 1, 74, 35]. Noting that the Prentice approach is based purely on statistical parameters and the study of associations between observable random variables, Joffe and Greene [34] suggested an alternative framework based on controlled direct and indirect causal parameters that assume experimental manipulation of the hypothesized surrogate, a framework also studied by Robins and Greenland [53] and Pearl [44]. While this controlled effects framework has major advantage to address questions about how interventions on the surrogate causally effect the clinical outcome, it is challenged by questions of conceivability of the causal target parameters in some settings [24] and by difficulties in justifying assumptions used to identify the causal parameters.

Observing that many early methods for assessing the Prentice criteria did not account for the fact that baseline predictors of both the surrogate and clinical outcome must be correctly controlled for, Frangakis and Rubin [18] introduced the principal stratification framework that studies how the clinical treatment effect varies over principal strata subgroups defined by the potential surrogate endpoints under each of the two treatment assignments. Many statistical methods papers in this framework have followed, including Gilbert and Hudgens [23], including Taylor, Wang, and Thiebaut [61], van der Weele [67], Li, Taylor, and Elliott [37], Huang, Gilbert, and Wolfson [30], and Gilbert et al. [22]. The meta-analysis framework studies the association of treatment effects on the surrogate outcome with treatment effects

on the clinical outcome [e.g., Daniels and Hughes [13], Buyse et al.[8], and Gail et al. [21]], with advantage that the treatment effects are causal effects based on the randomization and are estimable from standard assumptions.

VanderWeele [68] reviewed how these four frameworks relate to criteria for guaranteeing a consistent surrogate, and Gilbert et al. [22] studied relationships between principal stratification criteria and the Prentice definition. Except for a segment of the meta-analysis literature, there is quite limited surrogate endpoint evaluation literature on methods for applying and assessing the validity of a surrogate endpoint in a new trial for inferring the causal treatment effect in that trial without including clinical endpoint data [22]. The small size of this literature may be surprising given the centrality of this objective in biomedical applications. Pointing to this gap in the literature, Pearl and Bareinboim [46, 2] introduced the causal selection diagram approach, to estimation and testing of the clinical treatment effect in a new setting based on a surrogate and baseline covariates, which may be viewed as a fifth framework for surrogate endpoint evaluation.

Our newly proposed approach does not fit squarely into any of the five frameworks, thereby constituting a sixth framework that we name the optimal surrogate approach. It departs from the principal stratification and controlled effects frameworks, aligning more closely with the other three, in being based purely on statistical parameters that are estimable under the basic assumptions typically made in randomized clinical trials. In particular, it aligns with the Prentice framework by taking as its starting point the Prentice definition of a valid surrogate endpoint. In fact, the optimal surrogate is constructed to guarantee satisfaction of the Prentice definition, a unique advantage compared to previous approaches. Our approach also departs from previous approaches by defining the optimal surrogate as an unknown parameter, such that its predicted values are used as the surrogate endpoint. Because this estimated optimal surrogate is consistent under standard assumptions, in trials with large sample sizes it approximately satisfies the Prentice definition.

The optimal surrogate approach is related to Prentice's [50] operational criteria. First, the best optimal surrogate will have treatment and candidate surrogate separately highly predictive of the final outcome, similar to the first two Prentice criteria. Second, it posits a no direct effect criterion for licensing correct inferences on the clinical treatment effect in the new trial, which is a conditional mean version of Prentice's 'full mediation' criterion. Moreover, our approach departs from the Prentice criteria by applying both to settings where the studied surrogate varies in both treatment arms and to settings where it only varies in the active treatment arm, which is important given the many applications where the latter scenario attains [23, 22], whereas in contrast the Prentice approach only applies to the former scenario, e.g., Chan et al. [10] and Gilbert, Qin, and Self [25]. This is important because the latter scenario is quite common, for example in trials where the candidate surrogate is a biomarker response endpoint that is structurally negative/zero for all placebo/control group recipients ([23]).

The optimal surrogate approach is related to the meta-analysis framework by addressing the common objective of inference on the clinical treatment effect in a future study without collecting the clinical outcome in that study [21]. However, it tackles this objective based on a single (or few) efficacy trials plus transportability assumptions that are different from the 'extrapolation' assumptions needed via meta-analysis– meta-analysis based inference on the association of trial-level surrogate and clinical treatment effects estimated from a series of trials and the assumption that the series of trials forms a correct basis for extrapolating the clinical treatment effect to the new setting not included in the series. Finally, the optimal surrogate approach breaks new ground by treating the surrogate endpoint problem as a supervised statistical learning problem. While historically methods evaluate a pre-selected univariable or low-dimensional vector candidate surrogate, the optimal surrogate approach allows all collected baseline and intermediate response data to potentially contribute to the optimal surrogate, based on unbiased machine learning, and does not require parametric

modeling assumptions.

1.10 Generation of the SimCYD14 and SimCYD15 Dengue Vaccine Efficacy Trial Pseudo Data Sets

We generated two example dengue vaccine efficacy trial data sets that are similar to the real CYD14 and CYD15 dengue vaccine efficacy trial data sets [70, 9], except that instead of using case-cohort sampling designs as in the real trials, the generated data sets have all variables for all study participants. In addition, the simulated data sets simulate new data for all observations, introducing stochastic variability, to satisfy requirements from the sponsor of the CYD14 and CYD15 trials (Sanofi Pasteur) that this initial analysis does not disclose real data results. For each of the CYD14 and CYD15 trials, respectively, baseline and month 13 serotype neutralization titer measurements were made in a subcohort randomly sampled at study entry of approximately 10% and 20% of study participants. We refer to these neutralization titer measurements, obtained via the $PRNT_{50}$ plaque reduction neutralization test, as "serotype titers."

For study participants who completed the 25-month follow-up period free of the dengue disease primary study endpoint (the "controls"), we simulate data values based on these random samples. For dengue "cases" that had the primary dengue disease endpoint after month 13 and by month 25, month 13 serotype neutralization titer measurements were made in all cases, both within and outside the subcohort, which follows classical case-cohort sampling. However, baseline neutralization titer measurements were not made for cases outside of the subcohort, and we use the scheme described below for imputing these values for the simulated data sets. Additionally, due to the nature of the assay used to measure the baseline and month 13 serotype titers, titer levels below $log_{10}(10)$ were considered too low to quantify, and subsequently were censored at the value of $log_{10}(5)$. The imputation approach took this into account.

The process to impute missing baseline serotype titer values in dengue cases without

baseline serotype titer values for simCYD14 is outlined as follows:

1. For each serotype $k \in (1, 2, 3, 4)$, distributions for the baseline (B) and month 13 (M13) serotype titers were modeled:

$$\begin{aligned} S_k^B &\sim N(\mu_k^B, \left(\sigma_k^B\right)^2), \\ S_k^{M13} &\sim N(\mu_k^{M13}, \left(\sigma_k^{M13}\right)^2), \end{aligned} \tag{1.11}$$

   where μ_k^B, $\left(\sigma_k^B\right)^2$, μ_k^{M13}, $\left(\sigma_k^{M13}\right)^2$ were calculated as the sample means and variances of the baseline or month 13 serotype titer values for those who had data (for each serotype, respectively). These distributions [equation (1.11)] are biased slightly upwards as they used the censored data values [$log_{10}(5)$ for all real values ranging from 0 to $log_{10}(10)$]; however, for our purposes, this is a close enough approximation.

2. For all observations with censored baseline or month 13 data for a given serotype k, a random draw from the tail of these distributions (S_k^B and $S_k^{M13} \leq log_{10}(5)$) was imputed. This new variable (called S_k^{*B} or S_k^{*M13}) for each observation equaled either the observed above threshold value or, in cases that were censored, the newly imputed below threshold values. This approach simulates what the unmeasured (below threshold) values might have been had we been able to measure them.

3. Using cases with complete data (n=51 for simCYD14), we modeled the probability of an above threshold (*positive*) baseline titer value using a logistic regression model with an indicator predictor variable of whether or not the month 13 titer was positive, an interaction of that indicator and the actual or simulated (for censored data) titer, and the 3-category covariate age. A separate predictive model was fit for each of the four

serotype titer variables separately $k \in (1,2,3,4)$:

$$\begin{aligned} logit\left(P\left[S_k^B = +\right]\right) = \beta_0 + \beta_1 I_{\left[S_k^{M13}=+\right]} + \beta_2 I_{\left[S_k^{M13}=+\right]} S_k^{*M13} \\ + \beta_3 I_{Age\in(5,11]} + \beta_4 I_{Age\in(11,14]} \\ \text{for } k \in (1,2,3,4) \end{aligned} \tag{1.12}$$

These four models were used to output the predicted probabilities of *positive* ($> log_{10}(10)$) baseline serotype titers for each case without baseline data (n=193), done separately for each serotype $k \in (1,2,3,4)$.

4. Using the simulated baseline and month 13 titers (S_k^{*B} and S_k^{*M13}, see step 1), we created a regression model to predict baseline titers from the month 13 titers:

$$S_k^{*B} = \beta_0 + \beta_1 S_k^{*M13} + \beta_2 I_{Age\in(5,11]} + \beta_3 I_{Age\in(11,14]} \tag{1.13}$$

5. Using this model [equation (1.13)], we imputed the missing baseline data for cases. For those with predicted (pred) positive baseline titers, we applied this regression model, while for those with predicted negative baseline titers, we assigned the threshold value $log_{10}(5)$:

$$S_k^{B,impute} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 S_k^{*M13} + \hat{\beta}_2 I_{Age\in(5,11]} + \hat{\beta}_3 I_{Age\in(11,14]} + \epsilon_{k,i} & \text{if } P_{pred}\left[S_k^B = +\right] > 0.5 \\ log_{10}(5) & \text{if } P_{pred}\left[S_k^B = +\right] \leq 0.5 \end{cases} \tag{1.14}$$

where $\epsilon_{k,i}$ is a random draw from $\epsilon_k \sim N(0, \sigma_k^2)$, with its distribution estimated from the residuals of the models fit in equation (1.13). These models helped retain some of the correlation structure found between month 13 serotype values, baseline serotype values, and age.

6. The model in equation (1.13) allowed for some observations with $P_{pred}\left[S_k^B = +\right] > 0.5$ to be assigned an imputed serotype titer less than $log_{10}(5)$. To address this, after imputation any imputed serotype values below the threshold were censored at $log_{10}(5)$, ensuring our simulated distributions of serotype titers are similar to those of the real data.

After a complete simulated data set of dengue cases was generated by making imputations via equation 1.14, we bootstrapped from these observations a data set of size n=250 to finalize the set of simCYD14 data set case observations. To finalize the set of simCYD14 data set control observations, we bootstrapped from the approximately 2000 controls a sample of 10,000 control participants. This process preserves the observed dengue endpoint incidence rate of about 250 cases for every 10,000 individuals in the study population. This resulted in an analysis data set of 250 cases and 10,000 controls with a vaccination prevalence of approximately two-thirds, again approximating the real data set. Once the final imputed and bootstrapped data sets were constructed, graphical diagnostics were done to make sure there were no noticeable differences between the bootstrapped and the original study populations.

The same process was used to generate a simulated CYD15 data set, based on the original data set for the simCYD15 vaccine efficacy trial, with one difference in the models described above that the protocol-specified age categories 9–11 and 12–16 were used. The CYD15 trial had approximately twice the study size (approximately 20,000 subjects) and nearly twice the number of cases (415); therefore the simCYD15 data set for analysis was composed of 20,000 simulated controls and 500 simulated cases.

1.11 Additional Analyses of the SimCYD14 and SimCYD15 Dengue Vaccine Efficacy Trial Data Sets

Figures 1.5-1.8 display baseline dengue serotype neutralization titers (W) and month 13 dengue serotype neutralization titers (S) by protocol-specified age covariate categories (W) and the treatment category A (where $A = 1$ for vaccine and $A = 0$ for placebo). For both simCYD14 and simCYD15 it is apparent that older children tend to have higher serotype titers for all 4 serotypes than do younger children. Additionally for both studies, there does not appear to be any differences in baseline serotype titer distributions between the vaccine and placebo groups (as expected based on the randomization); however, there is an observable difference in the distributions of month 13 serotype titers between the vaccine and placebo groups, with higher month 13 titers seen on average for the vaccine group. This is expected given that one of the designed purposes of the vaccination is to generate serotype titer responses.

Figure 1.9 shows the univariate generalized additive model (gam) fits for the discrete super-learner model described in Table 2 of the main article, for study simCYD14.

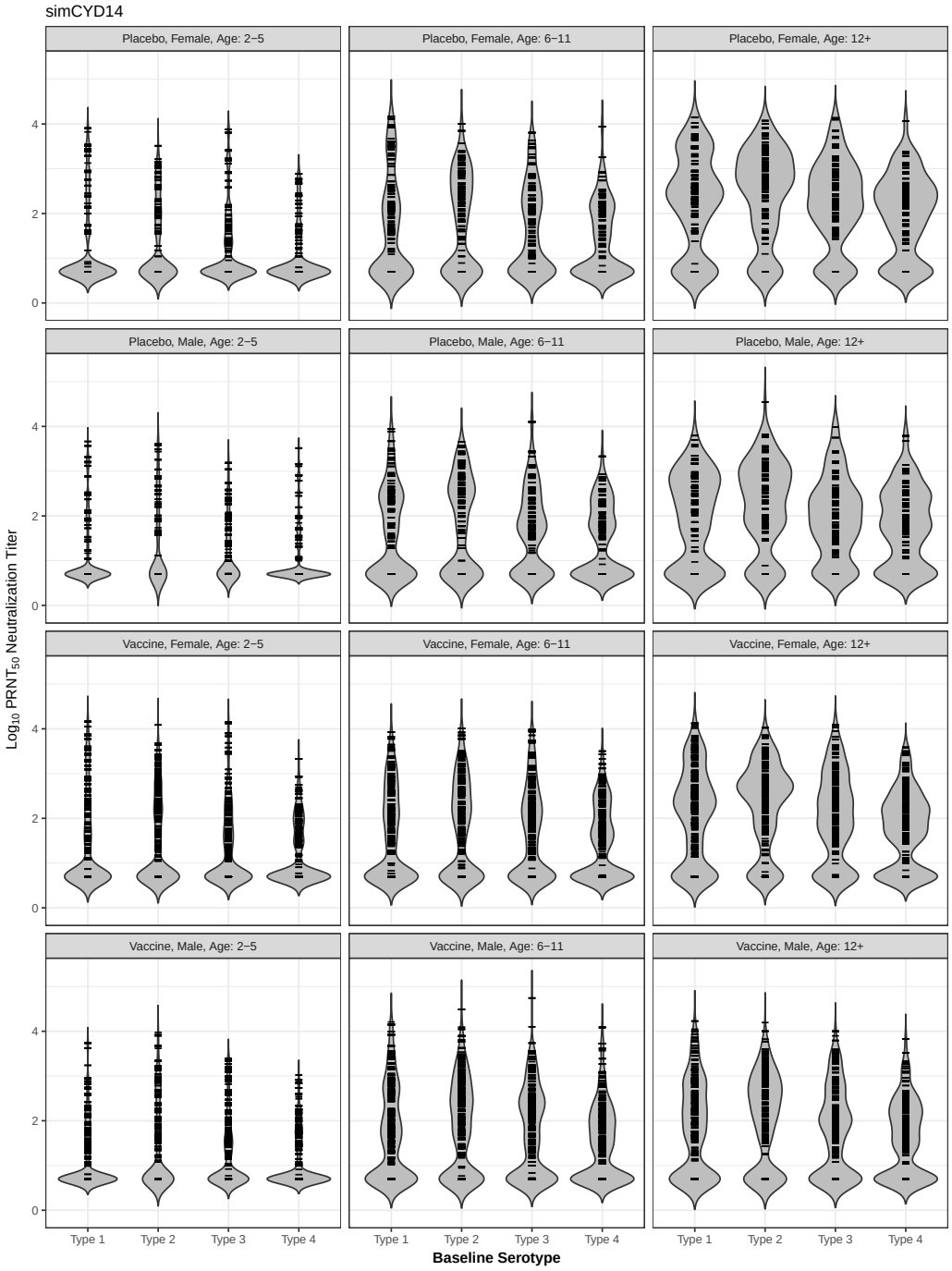

Figure 1.5: Distributions of log_{10} baseline dengue serotype neutralization titers ($PRNT_{50}$) for each of the 4 serotypes by sex and age categories for the simCYD14 trial

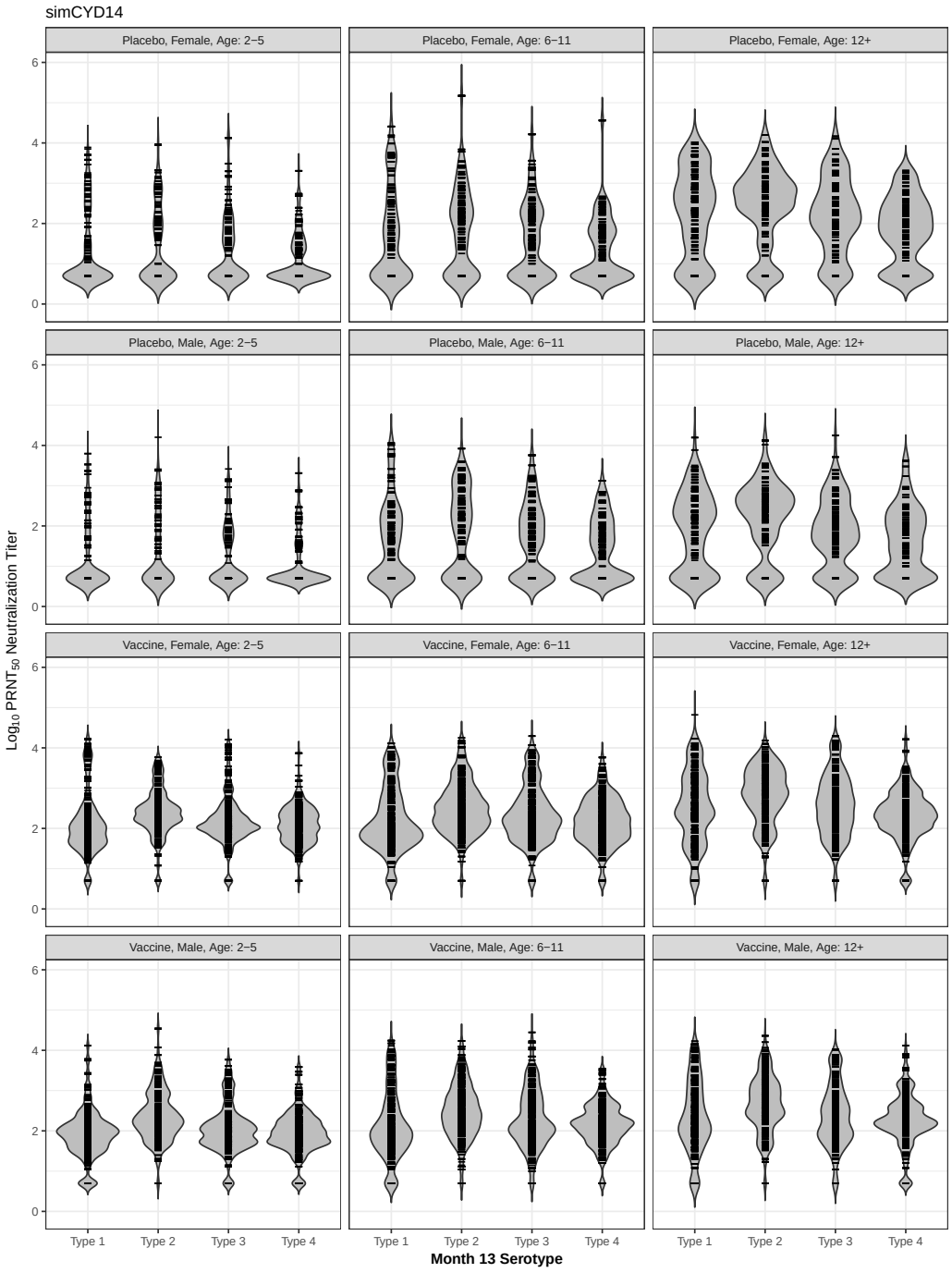

Figure 1.6: Distributions of log_{10} month 13 dengue serotype neutralization titers ($PRNT_{50}$) for each of the 4 serotypes by sex and age categories for the simCYD14 trial

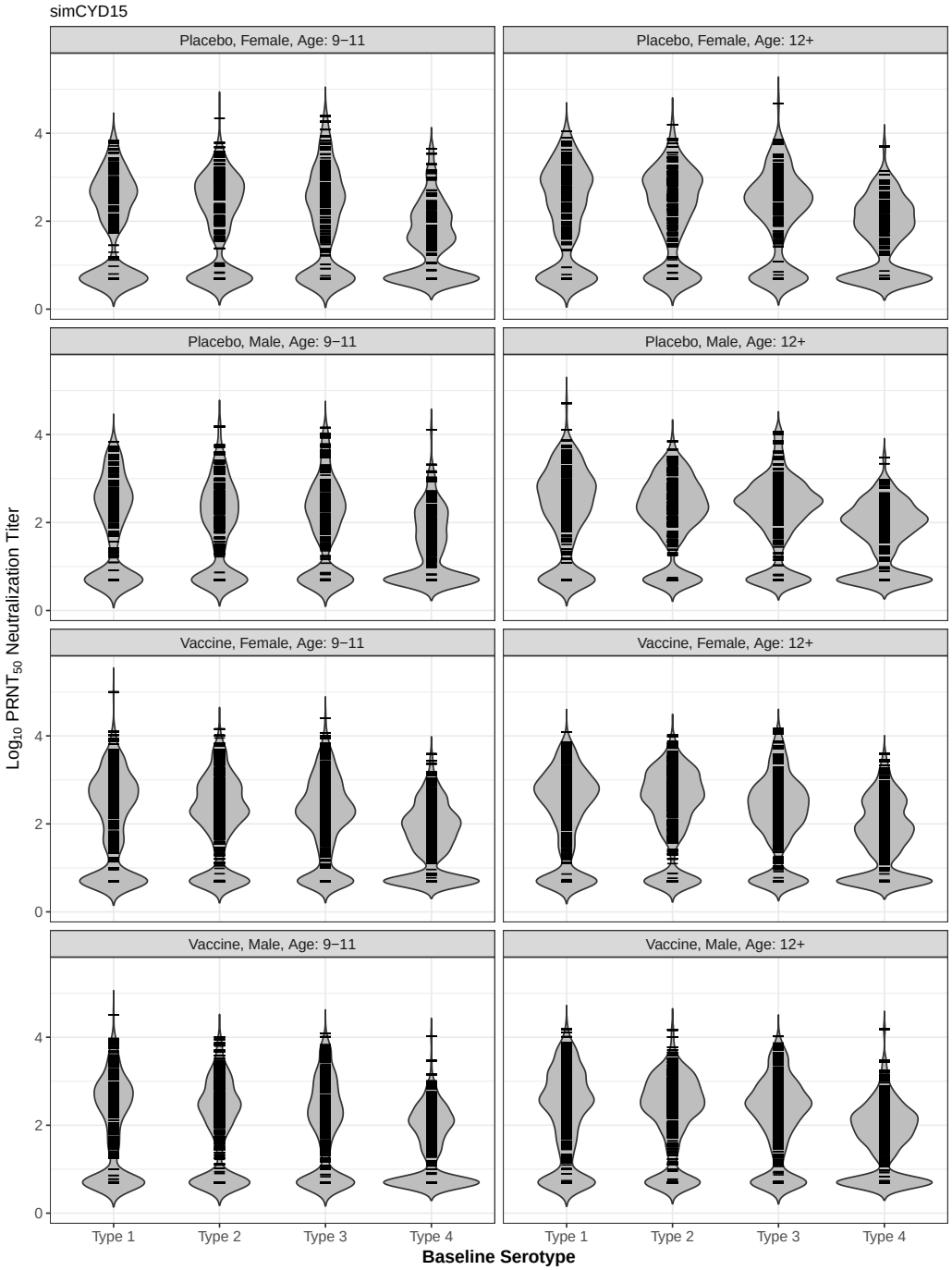

Figure 1.7: Distributions of log_{10} baseline dengue serotype neutralization titers ($PRNT_{50}$) for each of the 4 serotypes by sex and age categories for the simCYD15 trial

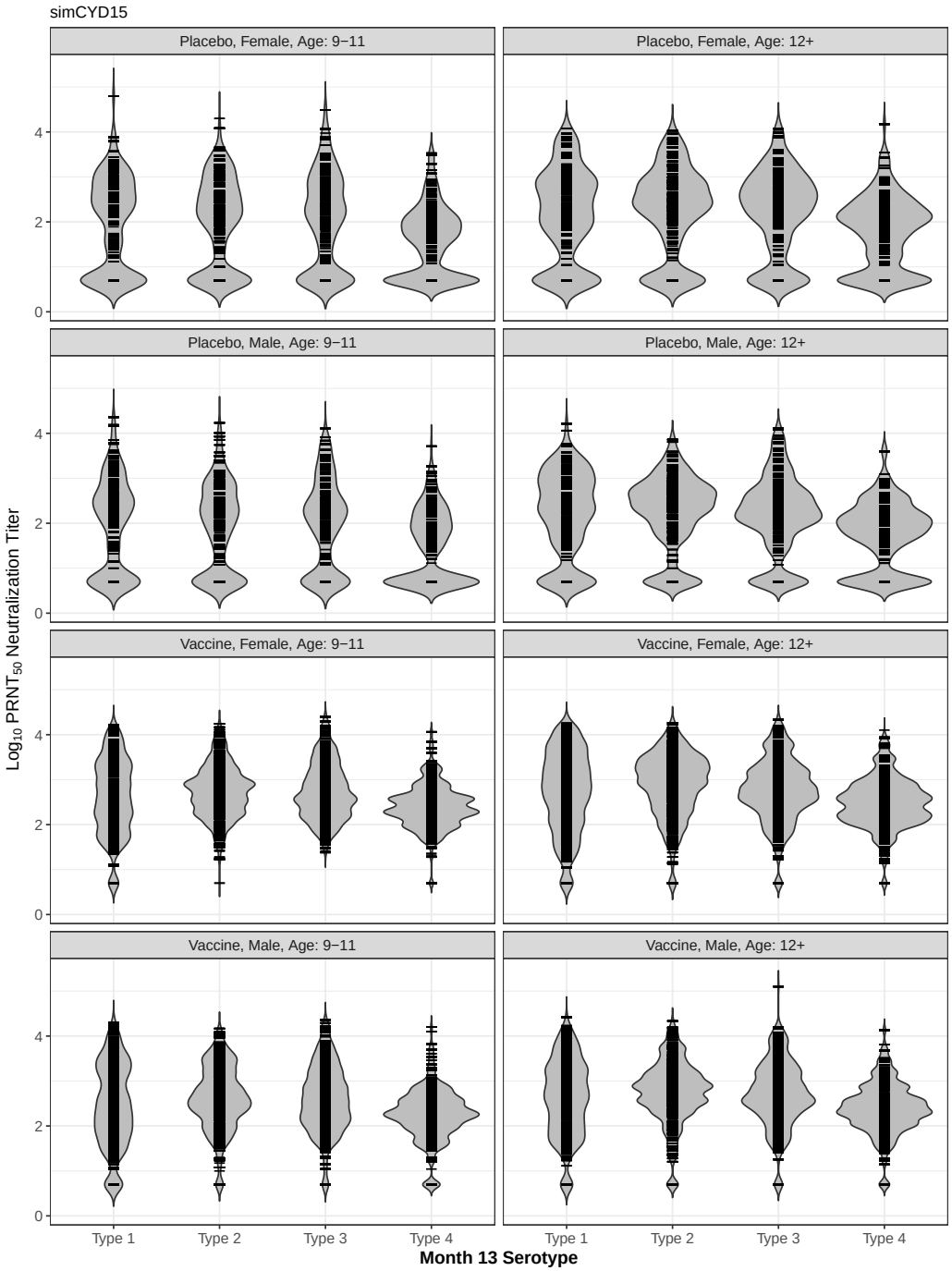

Figure 1.8: Distributions of log_{10} month 13 dengue serotype neutralization titers ($PRNT_{50}$) for each of the 4 serotypes by sex and age categories for the simCYD15 trial

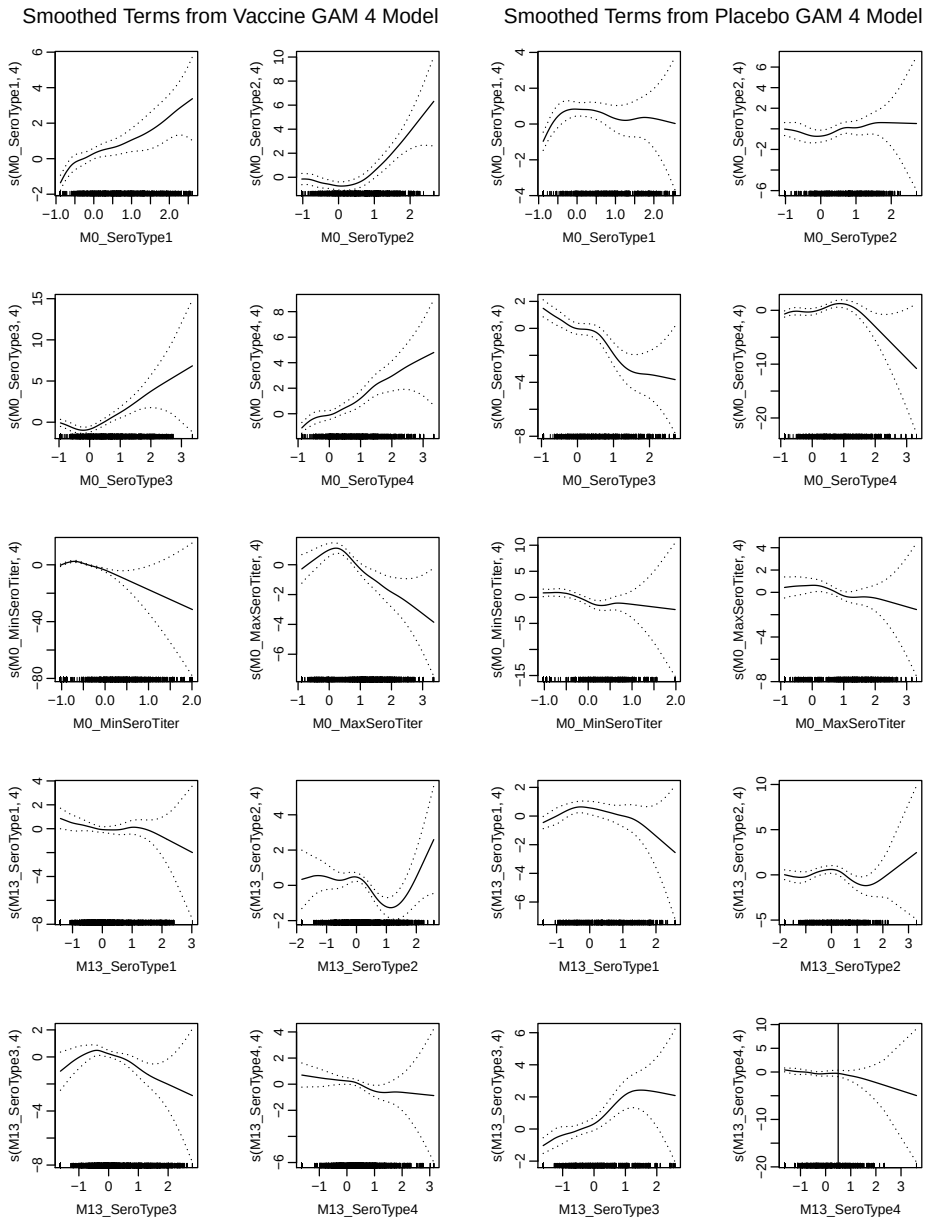

Figure 1.9: Smoothed term univariate relationships from the discrete super-learner for the vaccine and placebo group models fit to simCYD14 described in Table 2 of the main manuscript. Plots for each of the 10 smooth terms in the generalized additive models with 4 effective degrees of freedom are provided, with dotted lines representing a single standard error difference.

1.11.1 Diagnostic checking of the conditions of Theorem 2 for the dengue vaccine efficacy trials application

In Theorem 2 we established that the estimated optimal surrogate is still a valid surrogate in a new study under the following three conditions:

1. (Randomization) A remains randomized, conditional on W,

2. (Equal Conditional Means) the conditional mean of Y given (W, A, S) is the same as the conditional mean of Y^* given (W^*, A^*, S^*), and

3. (Contained Support) the support of (W^*, A^*, S^*) is contained in the support of (W, A, S).

We check these conditions treating simCYD14 as the original study and simCYD15 as the new study.

Condition 1 (Randomization) is met by the fact that both simCYD14 and simCYD15 randomized study participants to treatment (vaccine versus placebo).

Condition 2 (Equal Conditional Means) is explored in Figures 1.10 and 1.11, which display the difference between $\psi_n^{TMLE_{14}}(W, A, S)$ that was built using simCYD14 data and $\psi_{n*}^{TMLE_{15}}(W^*, A^*, S^*)$ that was built using simCYD15 data. For each fixed level of the observations in simCYD15, denoted by $(W^*, A^*, S^*) = (w, a, s)$, $\psi_n^{TMLE_{14}}(W, A, S) = E[Y|W = w, A = a, S = s]$ was calculated and subtracted from $\psi_{n*}^{TMLE_{15}}(W, A, S) = E[Y^*|W^* = w, A^* = a, S^* = s]$. Should the conditional mean of Y given (W, A, S) be identical to the conditional mean of Y^* given (W^*, A^*, S^*), these differences, $d(w, a, s) \equiv E[Y^*|W^* = w, A^* = a, S^* = s] - E[Y|W = w, A = a, S = s]$ should be close to zero for all observations in the simCYD15 data set. As can be seen in Figures 1.10 and 1.11, these differences generally cluster around zero and are centered around zero. The values are plotted by categories of age, sex, and treatment group, and are plotted against the baseline (Figure 1.10) and month 13 serotype values (Figure 1.10). For both baseline and month 13 serotype titer values, the

older age category of 12+ years for vaccinated individuals has a smaller spread of the differences $d(w, a, s)$ around 0, which was verified by the standard deviations for each category (results not shown). No other clear differences between covariate categories or along serotype titer values are apparent.

Condition 3 (Contained Support) requires the support of (W^*, A^*, S^*) to be contained in the support of (W, A, S). While the age range in simCYD14 was 2 to 14 years, the age range in simCYD15 was 9 to 16 years. In fitting models, the age variables were collapsed to "less than or equal to 5", "greater than 5 and less than or equal to 11", and "greater than 11" for simCYD14, and collapsed to "less than or equal to 11" and "greater than 11" for simCYD15 (the latter two age categories are specified in the CYD15 protocol). By year of age, the support of age for simCYD15 was not contained in that for simCYD14, since simCYD15 included 15 and 16 years old whereas the oldest participants in simCYD14 were 14 years old, thus violating the support assumption. However, there is evidence that children of age 15 and 16 only have slightly higher serotype titers than children of age 14, which suggests that this violation of the assumption may cause minor bias but is not expected to have a major impact on the results.

The support assumption trivially holds for the gender covariate, because both studies included male and female participants. In addition, all serotype titer variables (at baseline and month 13) had the same minimum values. The month 13 and baseline serotype-specific neutralization titers were also relatively similar between the two studies. The month 13 maximum serotype 3 neutralization titer values was 14% higher for simCYD15 than for simCYD14, but all other maximum month 13 titer values were smaller for simCYD15 when compared to simCYD14. For baseline titers, the serotype 1 maximum titer value was 18% greater and the serotype 4 maximum titer value was 2% greater for simCYD15 compared to the maximum titers for simCYD14. In sum, there are minor violations of the support assumptions that are expected to have a minor-to-moderate influence on the results.

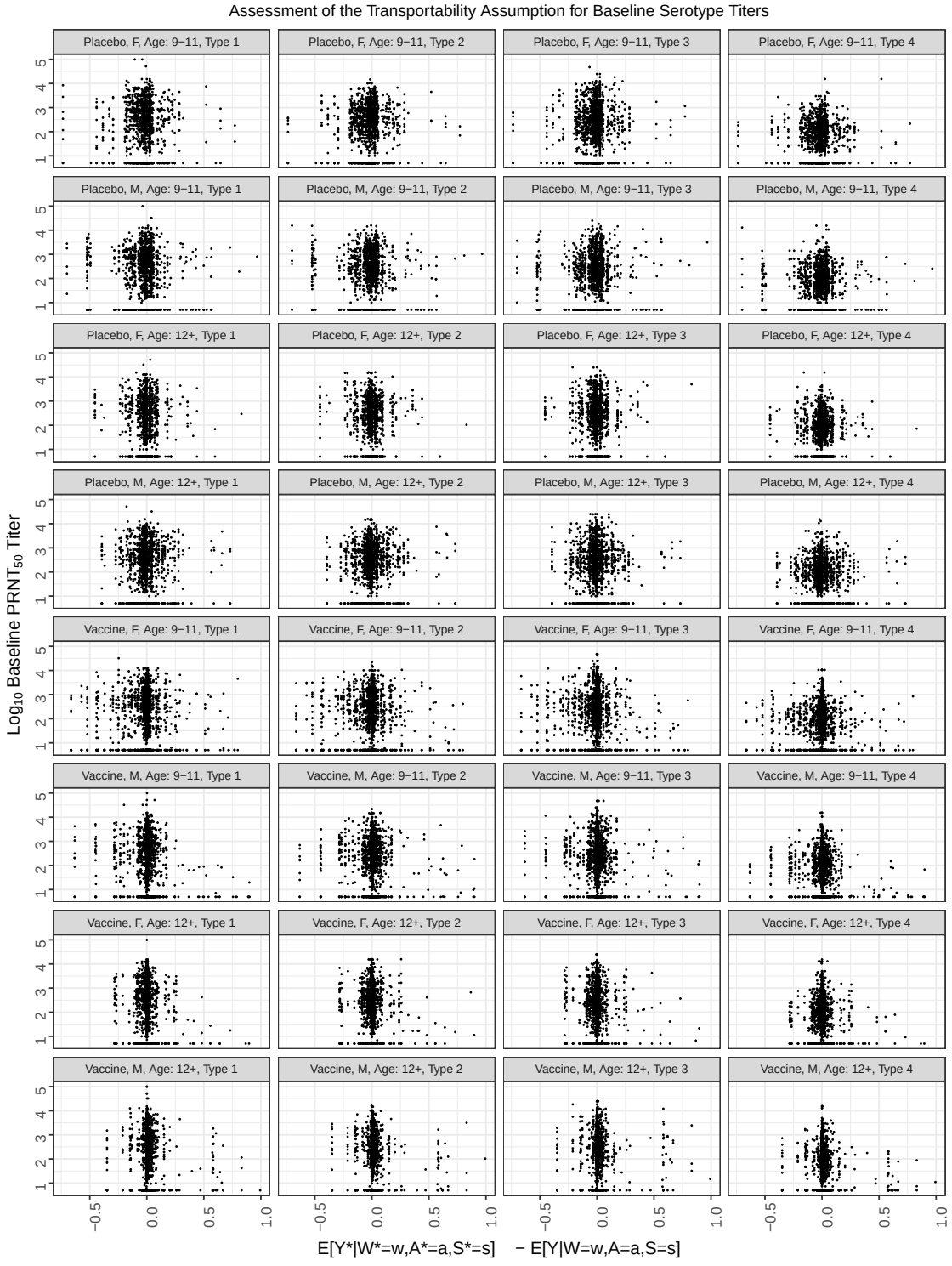

Figure 1.10: Diagnostics of the transportability assumption (Theorem 2): Plot of the differences (simCYD15 - simCYD14) in estimated optimal surrogate values for all observed values of simCYD15 participants, by covariate categories and baseline serotype titer values.

1.12 Considering Area Under the Receiver Operating Curve (AUC) Loss in the Application to Two Simulated Dengue Vaccine Efficacy Trials

In Section 1.6 we presented the application of estimating the optimal surrogate to two simulated dengue vaccine efficacy trials using SuperLearner with MSE loss. Due to the low incidence of our outcome of interest VCD (2-3%), other loss functions may be equally or more appropriate for estimating the optimal surrogate. In this section, we consider the application to the two simulated dengue trials using an AUC loss function for the SuperLearner instead of the MSE error used in Section 1.6.

The cross-validated MSE and risk values for the SuperLearner applied using MSE and AUC loss-based approaches are shown in Figures 1.12 and 1.13, respectively. Comparison of the two sets of results revealed that the two approaches yield similar results, with the top performers for both loss functions being generalized additive models and the super-learner or discrete super-learner. The resulting super-learners were used to define the estimated optimal surrogate for each analysis.

Figure 1.14 displays the empirical reverse cumulative distribution function for the estimated optimal surrogates for the vaccine and placebo groups ($\hat{\Psi}_n(W,1,S)$, $\hat{\Psi}_n(W,0,S)$) when fit in SuperLearner using MSE loss and AUC loss, using all input variables available. On average, cases (VCD) had a higher predicted probability of VCD (estimated optimal surrogate) than controls (no VCD) and the relationships are very similar across loss functions.

Using the AUC loss based superlearner, $\hat{\Psi}_n(W_i, A_i, S_i)$ was calculated for all observations and used to estimate the TMLE for vaccine efficacy ($\widehat{VE}_{TMLE}$). For simCYD14, $\widehat{VE}^{14}_{TMLE}$ was 52% (95% CI: 41%-66%), and the cross validated R^2 (CV-R^2) value for the estimated optimal surrogate (fit using AUC-based loss) for the vaccine and placebo groups were 0.17

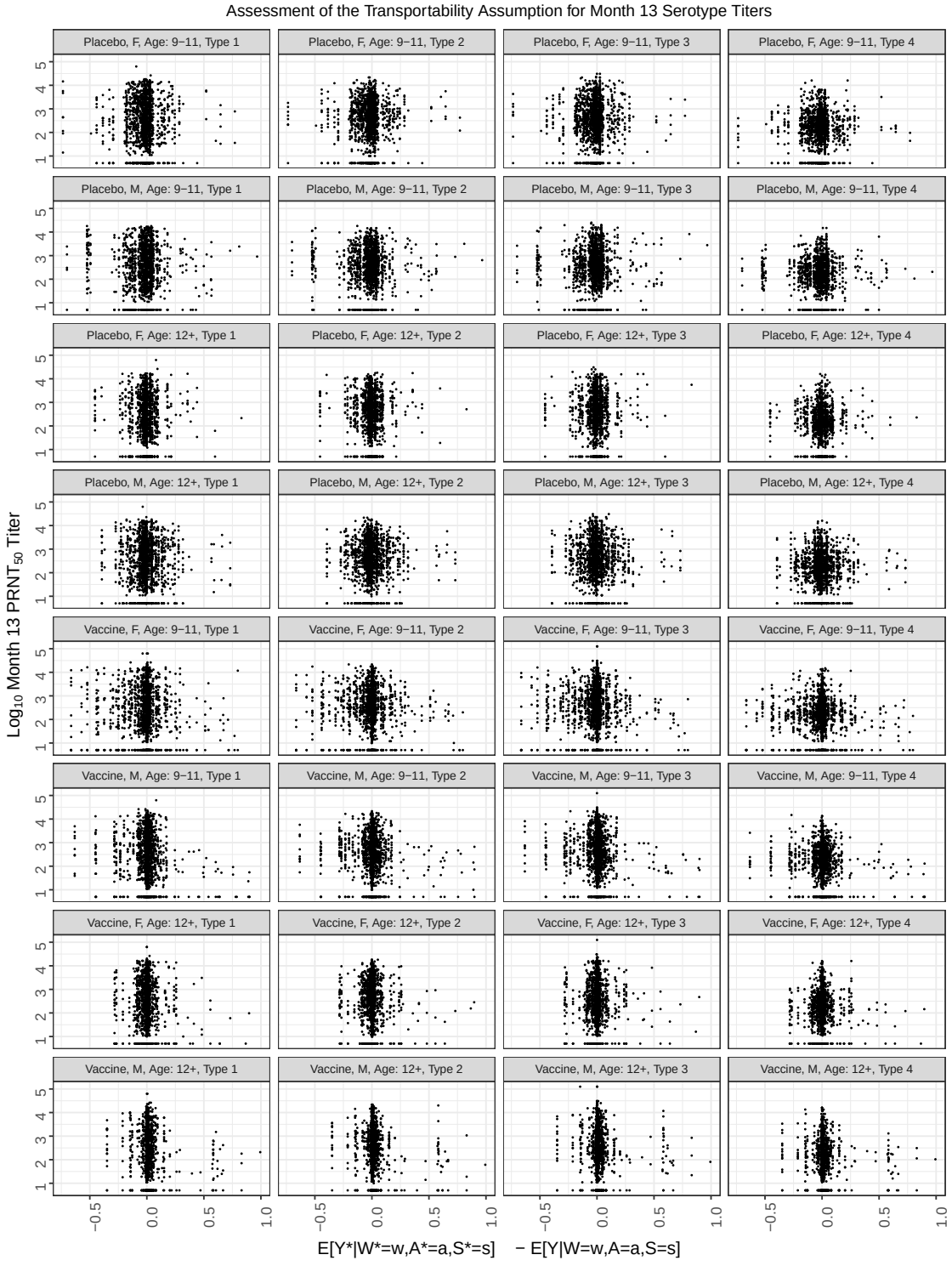

Figure 1.11: Diagnostics of the transportability assumption (Theorem 2): Plot of the differences (simCYD15 - simCYD14) in estimated optimal surrogate values for all observed values of simCYD15 participants, by covariate categories and month 13 serotype titer values.

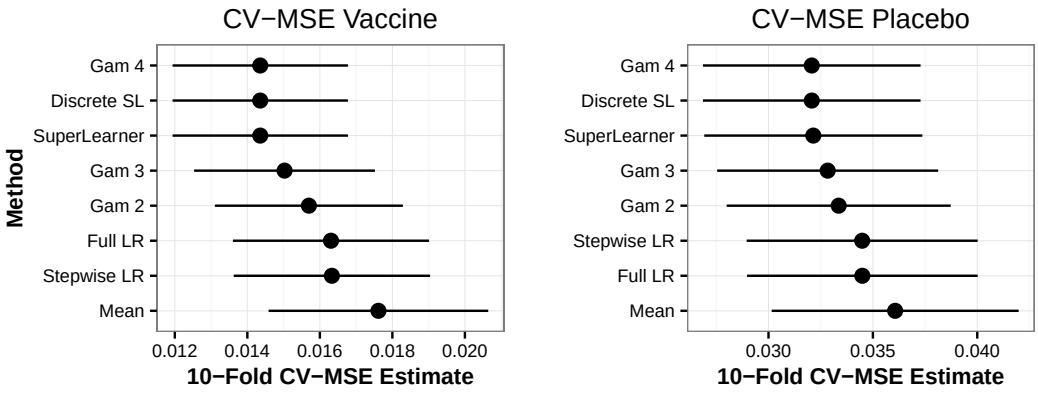

Figure 1.12: CV-MSE values with cross-validated 95% confidence intervals (CIs) for the vaccine and placebo groups. Lower values indicate a better fit to the data. SuperLearner, discrete SuperLearner, and the generalized additive model (degree 4) appear to do similarly well at predicting dengue outcome.

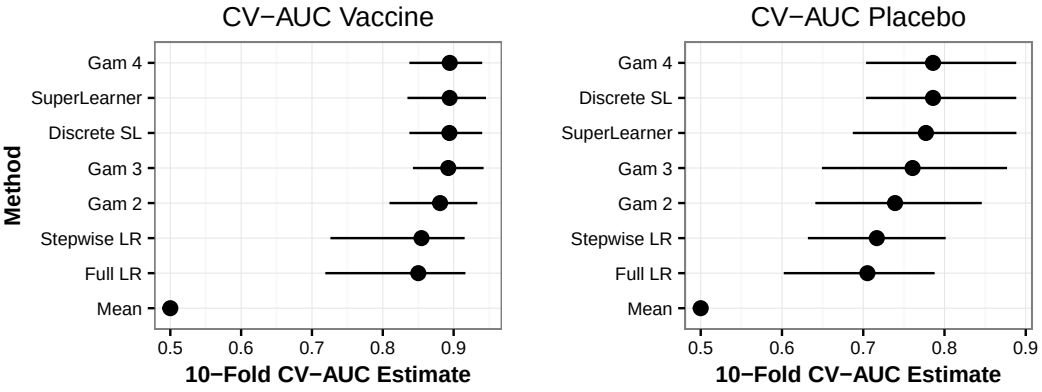

Figure 1.13: AUC loss values with cross-validated 95% CIs for the vaccine and placebo groups. Higher values indicate a better fit to the data, and we see that SuperLearner does the best at predicting dengue outcome, followed by 4th-degree generalized additive models for the vaccine and placebo groups, respectively.

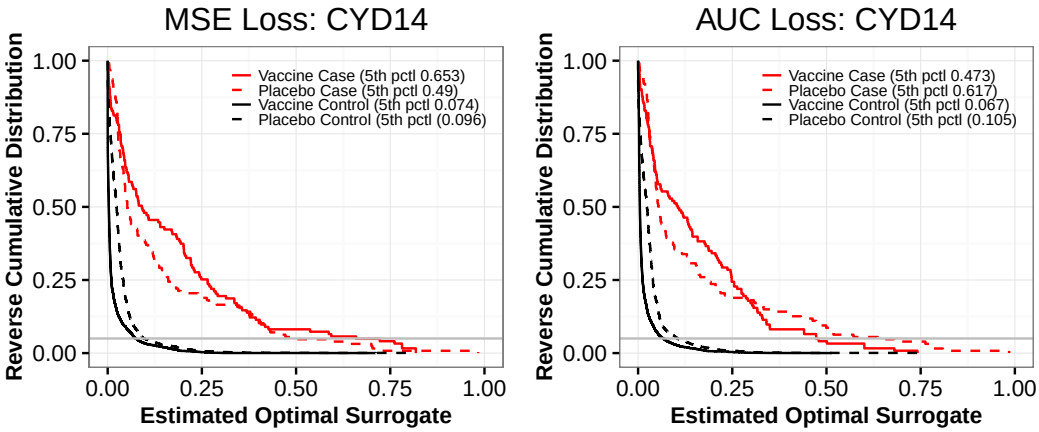

Figure 1.14: Reverse CDF function for the estimated optimal surrogate (SuperLearner) constructed using MSE loss and AUC based loss for the simCYD14 trial. We see that, on average, cases have higher predicted probability of VCD (optimal surrogate) values (red). Additionally, thresholds that correctly classify almost all controls also correctly classify most cases. Though the actual values of the estimated optimal surrogate differ between the loss functions, the relationship between thresholds set in the controls and classification rates in the cases is similar.

and 0.11, respectively. The cross validated AUC values for the estimated optimal surrogate for the vaccine and placebo groups were 0.89 (95% CI: 0.87, 0.92) and 0.78 (95% CI: 0.74, 0.82), respectively. Cross validated AUC was used to estimate AUC to account for the estimated optimal surrogate being fit on the simCYD14 data.

For comparison, the MSE loss based superlearner was also applied, $\hat{\Psi}_n(W_i, A_i, S_i)$ was calculated for all observations, and the TMLE for vaccine efficacy was once again estimated ($\widehat{VE}_{TMLE}$). For simCYD14, $\widehat{VE}^{14}_{TMLE}$ was estimated as 52% (95% CI: 42%-65%), and the cross validated R^2 (CV-R^2) value for the estimated optimal surrogate for the vaccine and placebo groups were 0.18 and 0.10, respectively. The cross validated AUC values for the estimated optimal surrogate for the vaccine and placebo groups when using MSE-loss based SuperLearning were 0.89 (95% CI 0.86, 0.92) and 0.77 (95% CI 0.73, 0.81), respectively. Cross validated AUC was used to estimate AUC to account for the estimated optimal surrogate

being fit on the simCYD14 data. Overall, the results were very similar between the two different loss function approaches.

1.12.1 Applying the estimated optimal surrogate to a new study: simCYD15 data

Once the estimated optimal surrogate $\hat{\Psi}_n(W, A, S)$ has been defined in a particular study, it is often desirable to use that newly defined surrogate as an endpoint in a future study, since the surrogate is often easier, more timely, and less expensive to measure than clinical endpoints. Moreover, the estimated optimal surrogate can be used to estimate VE in a new vaccine efficacy trial when the clinical endpoint is not available. We conducted this analysis for the simCYD15 vaccine trial, withholding the simCYD15 VCD clinical endpoint data and using the estimated optimal surrogate derived from simCYD14 plus estimated optimal surrogate measurements in simCYD15 to estimate VE in simCYD15. Importantly, however, VCD outcome data are available for simCYD15, enabling us to evaluate the performance of the estimated optimal surrogate in predicting VE in the new setting.

For this analysis, we applied the simCYD14 SuperLearner estimated optimal surrogate to the simCYD15 study participants, providing each subject with a predicted probability of disease and subsequently classifying the simCYD15 participants into cases and controls. In particular, after estimating $\hat{\Psi}_n(W, A, S)$ from the simCYD14 data, we calculated $\hat{\Psi}^*_n(W^*, A^*, S^*)$ for each observation in the new setting of simCYD15, and compared these fitted values to the known case and control outcomes Y. As shown in the reverse cumulative distribution function plots from this analysis, the estimated optimal surrogate still provides some power to distinguish between the cases and controls, with cases having higher average predicted probabilities of dengue (Figure 1.15). Using this estimated optimal surrogate in simCYD15 for the vaccine group, the threshold that correctly classified 95% of controls correctly classified 36% of cases; for the placebo group, the threshold that would correctly classified 95% of controls would correctly classified 24% of cases.

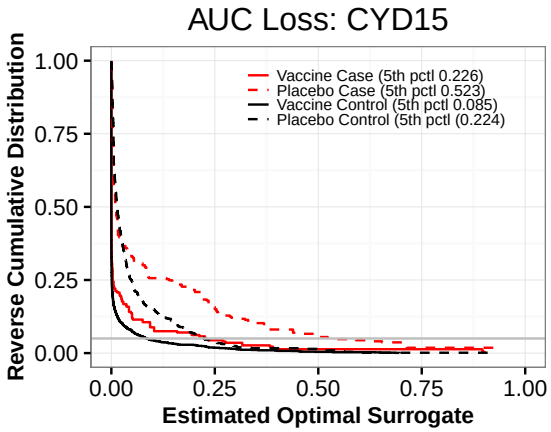

Figure 1.15: Reverse CDF function of the estimated optimal surrogate for simCYD15 vaccine and placebo recipients. The surrogate used here was the one fit from the simCYD14 trial using AUC loss. We see that, on average, simCYD15 cases again have higher predicted probability of VCD (estimated optimal surrogate) values. Additionally, thresholds that correctly classify most controls (no VCD) also correctly classify a large number of cases.

For simCYD15, the AUC for the estimated optimal surrogate was 0.60 (95% CI: 0.57-0.63) for the vaccinated group and 0.49 (95% CI: 0.45-0.53) for the placebo group. These findings indicate that an estimated optimal surrogate generated from data obtained in an earlier and similar, yet independent, clinical trial (simCYD14) could be used to obtain a reasonably accurate classifier of VCD in a second clinical trial, simCYD15. Our results also imply that the estimated optimal surrogate for the vaccinated group has stronger predictive ability than that for the placebo group, implying that the month 13 serotype titer values make a stronger contribution to predicting dengue in vaccine recipients than placebo recipients.

Based on the estimated optimal surrogate values $\hat{\Psi}_n^*(W^*, A^*, S^*)$ for simCYD15 vaccine and placebo recipients, we then calculated the TMLE for vaccine efficacy ($\widehat{VE}^{15}_{TMLE}$) for simCYD15 (which does not use dengue endpoints Y in simCYD15). We compared this estimate with that obtained from the simple empirical estimator of VE in simCYD15 based on the VCD endpoint data, ($\widehat{VE}^{15}_{Y}$), which was calculated as 1 - relative risk, where relative risk

was estimated from simCYD15 data as $\hat{E}[Y = 1|A = 1]/\hat{E}[Y = 1|A = 0]$. The results were $\widehat{VE}^{15}_{TMLE} = 68\%$ (95% CI: 58%-81%), and $\widehat{VE}^{15}_{Y} = 59\%$, (95% CI: 51%-65%), demonstrating that using the estimated optimal surrogate fit from simCYD14 on a similar but independent study, simCYD15, gives a reasonably similar estimate of VE compared to when VE is calculated using the actual clinical outcome data.

Chapter 2

OPTIMAL SURROGATE THEORY EXTENDED TO TWO PHASE STUDY DESIGNS THROUGH IPCW WITH SUPER LEARNING AND TMLE FOR CLASSIFICATION AND INFERENCE

Chapter 1 demonstrates the methodology for estimating an optimal surrogate in the context of complete baseline and intermediate time point data. Often in actual randomized trials, study design and data collection reflect a two-phase design in which intermediate timepoint data have been collected for only a subset of subjects (e.g., expensive biomarkers), frequently determined by values of the measured baseline covariates or outcome. This chapter will extend the optimal surrogate approach for complete data to two-phase data structures by using inverse probability of censoring weighted (IPCW) versions of Super Learner and TMLE, building on Rose and van der Laan [55].

2.1 Implementing the Methodology of Rose and van der Laan for Properly Weighting the TMLE Estimates for Two-Phase Designs

2.1.1 Description of A Targeted Maximum Likelihood Estimator for Two-Phase Designs as put forth by Sherri Rose and Mark J. van der Laan

Rose and van der Laan consider the application of appropriate weighting techniques to two-phase sampling designs, in particular, to studies in which a researcher takes a random sample from a target population of interest and then obtains measurements for each subject in this first phase, and then follows that with a second phase in which a subsample of that original sample is drawn, based on the information obtained at the first phase. At the second phase

additional information is collected for only the subsample, and this phase-two subsample is used as the analysis data set. This type of data structure can be thought of as a missing data problem, in which data is missing for the full-data structure based on the second phase (subsample) structure. Rose and van der Laan proposed a methodology using inverse probability of censoring weighted targeted maximum likelihood estimator (IPCW-TMLE) for these types of two-phase sampling designs.

A brief summary of their methodology begins by representing the observed data structure for a sampled subject as $O = (V, \Delta, \Delta X)$, where V represents the information collected in the first phase on all sampled subjects, Δ represents an indicator of whether or not the subject was selected for the second phase and X represents the full-data structure collected in the second-phase of the study. Our full sample could then be represented as n i.i.d. copies of $O_1 \dots O_n$ of O. Additionally, in two-phase studies an outcome of interest Y would be included in V, and subjects could be sampled conditional on Y. If $P_{X,0}$ is the true probability distribution of X, and $\mathcal{M}^F$ is a statistical model for $P_{X,0}$, we can define our target parameter of interest as $\Psi^F(P_{X,0}) : \mathcal{M}^F \to \mathbb{R}^d$. Then the parameter of the true probability distribution of X would be $\psi_0^F = \Psi^F(P_{X,0})$, and the efficient influence curve of Ψ^F at a full-data distribution P_X would be denoted as $D^F(P_X)$.

We can write the conditional probability of the distribution of Δ given X as $g_{\Delta,0}(\delta|X) = P_{X,0}(\Delta = \delta|X)$. The probability of selection for the second phase given X is an important quantity for the IPCW-TMLE approach and, for notational convenience, is often written as $\Pi_0(X) \equiv g_{\Delta,0}(1|X)$. If we assume missing at random (MAR), which can be arranged to hold by the choice of sampling design, we then have that Δ is independent of X given V ($g_{\Delta,0}(\delta|X) = g_{\Delta,0}(\delta|V)$) and thus, with some abuse of notation, $\Pi_0(X) = \Pi_0(V)$. This value $\Pi_0(V)$, the probability of sampling given V, will contribute to the weight in future calculations and can be empirically estimated from the data $(\Delta_i, V_i), i = 1, \dots, n$. The efficient influence curve of $\Psi^F(P_{X,0})$ is an identifiable parameter of P_0 and is denoted by

$D^*(P_0) = D^*(P_{X,0}, \Pi_0)$.

Rose and van der Laan show that computing an initial IPCW-loss based estimator would be based on the IPCW-loss function that is weighted using an estimate of $\Pi_0(V)$, which we write as $\Pi_n(V)$. First, our loss function now includes the weight estimate,

$$L(P_X)(O) \equiv \frac{\Delta}{\Pi_n(V)} L^F(P_X)(X)$$

and therefore leads to using a weighted ϵ for the TMLE updating step. When applying the IPCW-TMLE approach, for $k = 1, \ldots, K$ one computes the amount of fluctuation in a way that now includes $\Pi_n(V)$:

$$\begin{aligned} \epsilon_n^k &= \arg\min_{\epsilon} P_n L(P_{X,n}^{k-1}(\epsilon)) \\ &= \arg\min_{\epsilon} \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\Pi_n(V_i)} L^F(P_{X,n}^{k-1}(\epsilon))(X_i) \end{aligned}$$

This updating process is iterated until convergence and the final update is denoted with $P_{X,n}^*$ which is called the IPCW-TMLE of $P_{X,0}$. From there, $\psi_n^{TMLE} = \psi_n(P_{X,n}^*)$ can be calculated, along with appropriate Wald-based $(1 - \alpha)$ confidence intervals.

When we apply SuperLearner or any other methodology to estimate our P_0-optimal surrogate from data following a two-phase design, the IPCW weights are necessary for valid initial estimates, and are used in the subsequent IPCW-TMLE estimate methodology for valid inference. This means that we estimate $E_0(Y|W, A, S)$ with a minimizer of the risk of our weighted loss function ($L(\psi)(O) \equiv \frac{\Delta}{\Pi_n(V)} L(\psi)$):

$$\psi_n = \arg\min_{\psi} P_n L(\psi) = \arg\min_{\psi} E_0 \frac{\Delta}{\Pi_n(V)} L(\psi)$$

.

In the optimal surrogate framework, we defined the P_0-optimal surrogate as

$$\psi_0^F = \arg\min_{\psi \in \mathbf{\Psi}} MSE_{P_{X,0}}(\psi).$$

So now, in the context of a 2-phase design, the weights contribute to the estimation of $E_0(Y|W, A, S)$, and thus our optimal surrogate is the solution of

$$\psi_0^F = \arg\min_{\psi \in \mathbf{\Psi}} E_{P_0} \frac{\Delta}{\Pi_n(V)}(Y - \psi(W, A, S))^2.$$

In our optimal surrogate framework, when targeting our estimate in a two-phase study, we can consider the submodel $\text{Logit}\psi_n(\epsilon) = \text{Logit}\psi_n + \epsilon H_{g_n}$ where $H_{g_n}(W, A, S) = (2A - 1)/[\Pi_n(V)g_n(A|W)]$ and g_n is an estimator of g_0. The covariates W, outcome Y, and treatment A are measured at the first phase and are thus contained within V as defined earlier. Those intermediate measurements made at phase-two (S) would be included in X as defined by Rose and van der Laan. We now let $\epsilon_n = \arg\min_\epsilon P_n L(\psi_n(\epsilon))$, with a loss function for our estimated optimal surrogate:

$$L(\psi)(O) \equiv -\frac{\Delta}{\Pi_n(V)} \log \psi(W, A, S)^Y (1 - \psi(W, A, S))^{1-Y},$$

a weighted version of the log-likelihood loss function or perhaps

$$L(\psi)(O) \equiv \frac{\Delta}{\Pi_n(V)} \{g(A|W)(Y - \psi(A, W, S))^2\},$$

a weighted version of mean squared error loss. When we let $\psi_n^{TMLE} = \psi_n(\epsilon_n)$ be the corresponding targeted estimator of ψ, we note that

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\Pi_n(V_i)} H_{g_n}(W_i, A_i)(Y_i - \psi_n^{TMLE}(W_i, A_i, S_i)).$$

For estimating our target parameters, the general formula to estimate $E[Y_1|W,A,S]$, $E[Y_0|W,A,S]$ and VE without weights using the TMLE method is

$$\begin{aligned}\hat{E}[Y_1|W,A,S]^{TMLE} &= \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(A=1,W_i,S_i)\\ \hat{E}[Y_0|W,A,S]^{TMLE} &= \frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(A=0,W_i,S_i)\\ \hat{VE}^{TMLE} &= 1-\frac{\frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(A=1,W_i,S_i)}{\frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n(A=0,W_i,S_i)}\end{aligned} \tag{2.1}$$

and now, with the weights to account for two-phase sampling, would be

$$\begin{aligned}\hat{E}[Y_1|W,A,S]^{TMLE}_{\Pi_n(V)} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\Pi_n(V_i)}\bar{Q}_n(A=1,W_i,S_i)\\ \hat{E}[Y_0|W,A,S]^{TMLE}_{\Pi_n(V)} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\Pi_n(V_i)}\bar{Q}_n(A=0,W_i,S_i)\\ \hat{VE}^{TMLE}_{\Pi_n(V)} &= 1-\frac{\hat{E}[Y_1|W,A,S]^{TMLE}_{\Pi_n(V)}}{\hat{E}[Y_0|W,A,S]^{TMLE}_{\Pi_n(V)}}.\end{aligned} \tag{2.2}$$

2.2 Application of IPCW-TMLE for Targeted Estimation of a Desired Parameter

As described above, we start with a data structure of $O=(V,\Delta,\Delta X)\sim P_0$ and our desired target parameter $\Psi(P_0)$. For illustration purposes we will designate $\Psi(P_0)=1-E_{W,0}[E_0(Y|A=1,W)/E_0(Y|A=0,W)]$, which represents the causal vaccine efficacy under the causal assumptions. Once again, $\Pi_0(X)\equiv g_{\Delta,0}(1|X)$, the probability of sampling, and we assume missing at random (MAR), such that Δ is independent of X given V, eg $g_{\Delta,0}(\delta|X)=g_{\Delta,0}(\delta|V)$, and thus $\Pi_0(X)=\Pi_0(V)$, with $\Pi_n(V)$ denoting the empirical estimate of $\Pi_0(V)$.

Therefore, our substitution TMLEs for vaccine efficacy (VE) can be written as

$$\psi^{VE}_{n,\Pi_n(V)} = \Psi(Q_n) = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\Pi_n(V_i)} \bar{Q}_n(1, W_i, S_i)}{\frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i}{\Pi_n(V_i)} \bar{Q}_n(0, W_i, S_i)}$$

with $Q_n = (\bar{Q}_n, Q_{W,n})$ and $Q_{W,n}$ as the empirical distribution for the marginal distribution of W. These values are estimated initially using SuperLearner. The initial estimate of $\bar{Q}_0$, denoted with subscript n as $\bar{Q}^0_n$ is then updated with TMLE iteratively to target the parameter of interest. In this example the treatment mechanism is $g_0 = P_0(A|W)$, which can be specified by the randomization protocol, but can also be estimated using SuperLearner with the estimate denoted as g_n. As stated in Moore and van der Laan's work [41] as well as in van der Laan and Robins [64], efficiency increases by estimating g_0 from the data if $Q(A, W)$ is not correctly specified. "The TMLE is still consistent when Q(A, W) is mis-specified; however, we can gain efficiency when estimating the treatment mechanism in such a case" [41]. Therefore, it is desirable to use SuperLearner to obtain g_n even when a treatment assignment protocol is available. Even though this estimate can be simply determined using a logistic regression model, a SuperLearner fit will allow for a potentially more precise estimate.

Once the estimate g_n is obtained, a working submodel can be applied for iteratively fitting the IPCW-TMLE. For a binary treatment, a logistic regression that regresses the covariates W on the treatment assignment A is a common approach. In the IPCW-TMLE framework, this regression would be weighted using the IPCW observation weights. For estimating our parameter of interest, $\psi^{VE}_{n,\Pi_n(V)}$, the "clever covariate" used in the logistic working submodel, $h(A, W)^{RR}_{\Pi_n(V)}$,would now utilize weight-adjusted $\hat{g}_0(1|W)$, $\hat{g}_0(0|W)$, μ_1 and μ_0:

$$h(A, W)^{RR}_{\Pi_n(V)} = \frac{\Delta}{\Pi_n(V)} \left[\frac{1}{\hat{\mu}_1} \frac{I(A = 1)}{\hat{g}_0(1|W, S)} - \frac{1}{\hat{\mu}_0} \frac{I(A = 0)}{\hat{g}_0(0|W, S)} \right]$$

where

$$\hat{\mu}_1 = \hat{E}[Y_1|W,S]^{TMLE}_{\Pi_n(V)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\Pi_n(V_i)}\bar{Q}^1_n(A=1,W_i,S_i)$$
$$\hat{\mu}_0 = \hat{E}[Y_0|W,S]^{TMLE}_{\Pi_n(V)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\Pi_n(V_i)}\bar{Q}^1_n(A=0,W_i,S_i)$$

And, our submodel ($\text{Logit}\psi_n(\epsilon) = \text{Logit}\psi_n + \epsilon H_{g_n}$) is now a weighted logistic regression based on the observation level IPCW weights.

The updated estimate of $\bar{Q}_0$ is denoted by $\bar{Q}^1_n$, with subsequent iterations of updated denoted by increasing superscripts (e.g. $\bar{Q}^2_n$, $\bar{Q}^3_n$. . .). In the case of these target parameters, only a single updating step is needed to obtain convergence [65], and thus $\bar{Q}^1_n$ can often be used for the plug-in targeted estimators described in Equation 2.2.

2.3 Influence Curves for Variance Estimates

For valid inference, estimates of the variance can be obtained by using the respective influence curves. On page 528 and 529 of van der Laan and Rose's book [65] the equation for the influence function IC_1 for $\psi(1) = E[Y_1|W,S]$ is given by $\bar{Q}_{1,0}(W) - \psi_0(1) + IC'_1(0)$ where

$$IC'_1 = \frac{I(A=1)}{g_0(1|W)}(Y - E_0(Y|A=1,W,S))$$

Therefore, the influence curve IC_1 for the treatment (vaccine) group mean ($E[Y_1|W,S]^{TMLE}$), when no weighting is applied, is equal to

$$IC_1 = \frac{I(A=1)}{g_0(1|W)}(Y - E_0(Y|A=1,W,S)) + \bar{Q}_{1,0}(W) - \psi_0(1)$$

and when weighting is applied becomes

$$\begin{aligned} IC_{1,\Pi_n(V)} &= \frac{\Delta}{\Pi_n(V)}\left(\frac{I(A=1)}{g_0(1|W)}(Y - E_0(Y|A=1,W,S)) + \bar{Q}_{1,0}(W) - \psi_0(1)\right) \\ &= \frac{\Delta}{\Pi_n(V)}\left(\frac{I(A=1)}{g_0(1|W)}(Y - \bar{Q}_{1,0}(W)) + \bar{Q}_{1,0}(W) - \psi_0(1)\right) \end{aligned}$$

where $\Pi_n(V)$ indicate the empirically estimated case-control weights and $\bar{Q}_{1,0}(W)$ indicates the TMLE adjusted predicted values.

And likewise, the weighted version for the control (placebo) group mean ($E[Y_0]^{TMLE}_{\Pi_n(V)}$) is

$$IC_{0,\Pi_n(V)} = \frac{\Delta}{\Pi_n(V)}\left(\frac{I(A=0)}{g_0(0|W)}(Y - \bar{Q}_{0,0}(W)) + \bar{Q}_{0,0}(W) - \psi_0(0)\right).$$

To calculate a valid estimate of the 95% confidence interval for $VE^{TMLE}_{\Pi(V)}$ one simply applies the delta method to estimate first the variance for the log relative risk, and thus estimate the 95% confidence interval for the log relative risk. A 95% confidence interval from there follows by transforming the 95% confidence interval for log relative risk back to the relative risk scale and then subtracting it from one to obtain the confidence interval for vaccine efficacy.

For the TMLEs for the individual treatment group means ($a \in \{0,1\}$), the 95% confidence intervals can be calculated as

$$\hat{E}[Y_a|A=a,W,S]^{TMLE}_{\Pi_n(V)} \pm 1.96 * \sqrt{\sum_{i=1}^{n} IC^2_{a,\Pi_n(V)}(O)/n}.$$

For the TMLEs for vaccine efficacy, one works with TMLE estimates of the log relative risk. After first calculating the TMLE and influence curves for the log relative risk, the

variance is then estimated via the delta method. Using estimates

$$\hat{\mu}^{TMLE}_{1,\Pi_n(V)} = \hat{E}[Y_1|W,S]^{TMLE}_{\Pi_n(V)}$$
$$\hat{\mu}^{TMLE}_{0,\Pi_n(V)} = \hat{E}[Y_0|W,S]^{TMLE}_{\Pi_n(V)}$$

one can calculate estimates

$$\hat{VE}^{TMLE} = 1 - \hat{RR}^{TMLE} = 1 - \hat{\mu}^{TMLE}_{1,\Pi_n(V)}/\hat{\mu}^{TMLE}_{0,\Pi_n(V)}$$
$$log(\hat{RR}^{TMLE}) = \log\left[\hat{\mu}^{TMLE}_{1,\Pi_n(V)}/\hat{\mu}^{TMLE}_{0,\Pi_n(V)}\right]$$
$$SE(log(\hat{RR}^{TMLE})) = \sqrt{\frac{1}{(\hat{\mu}^{TMLE}_{1,\Pi_n(V)})^2}Var\left(\hat{\mu}^{TMLE}_{1,\Pi_n(V)}\right)/n + \frac{1}{(\hat{\mu}^{TMLE}_{0,\Pi_n(V)})^2}Var\left(\hat{\mu}^{TMLE}_{0,\Pi_n(V)}\right)/n}.$$

From the above values the 95% confidence interval for the $\log(RR)$ can be written as

$$\log(RR^{TMLE}) \pm 1.96 * SE(\log(\hat{RR}^{TMLE}))$$

and therefore 95% confidence interval for VE as

$$1 - exp\left[\log(RR^{TMLE}) \pm 1.96 * SE(\log(\hat{RR}^{TMLE}))\right]$$

2.3.1 Code for weighted TMLE

Currently, the *tmle* package available does not directly handle observation level weights, and thus, to apply IPCW weighting for estimating a TMLE hand coding is necessary. Below is example code implementing weighted TMLE for Additive Treatment Effect (ATE) using a simple data set up. We can write our ATE parameter of interest as $\psi^{ATE}_{n,\Pi_n(V)} = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\Pi_n(V_i)}[\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)]$ and would utilize weight-adjusted $\hat{g}_0(1|W)$, $\hat{g}_0(0|W)$, $\hat{\mu}_1$ and $\hat{\mu}_0$ as well as the appropriate ATE clever covariate, $h(A,W,S)^{ATE}_{\Pi_n(V)} = \left[\frac{I(A=1)}{\hat{g}_0(1|W)} - \frac{I(A=0)}{\hat{g}_0(0|W)}\right]$.

Implementing the estimation of this ATE TMLE can be done using the simple code below.

Basic Weighted TMLE Function for Additive Treatment Effect (ATE)

```
weighted.tmle.ate <- function(
      w,  # the matrix of covariates considered; includes s
      a,  # a vector of treatment assignment
      y,  # a vector of outcomes
      Q.SL.library, # SuperLearner library for estimating Q
      g.SL.library,  # SuperLearner library for estimating g
      wgts,   #  the observation level weights
      max.wgt=Inf){
```

Set n equal to the number of observations

```
      n <- nrow(w)
```

Create matrix of actual covariates with assigned treatment, as well as matrices of covariates with either A = 1 or A = 0.

```
      WA <- WA0 <- WA1<-data.frame(w,A=a)
      WA0$A <- 0 # counterfactual for control
      WA1$A <- 1 # counterfacutal for case
```

Call SuperLearner to fit $\bar{Q}_0$ and save the initial estimate of the predicted values.

```
      require(SuperLearner)
      temp <- SuperLearner(y,data.frame(w,A=a), SL.library=Q.SL.library,
      newX=rbind(WA,WA0,WA1), obsWeights=wgts,
      family=binomial, cvControl=list(V=10))
```

```
Qbar.ests <- temp$SL.predict[,1]
rm(temp)
```

Estimate and save $g(A, W)$ using SuperLearner.

```
temp <- SuperLearner(a,w,SL.library=g.SL.library,
newX=w, obsWeights=wgts, family=binomial,
cvControl=list(V=10))}
g.ests <- temp$SL.predict[,1]
rm(temp)
```

Assign g_n based on the observed treatment group A. Adjust the indicator coding for A for convenience.

```
g.ests[a==0] <- 1-g.ests[a==0]
a.ind <- 2*a-1
```

Calculate the offset ($logit[\bar{Q}_n(A, W, S)]$) and ϵ from this equation:

$$logit\mu(A, W, S) = logit[\bar{Q}_n(A, W, S)] + \epsilon H(A, \Delta, W, S).$$

where

$$H(A, \Delta, W, S) = \frac{\Delta}{\Pi_n(V)} \frac{(2A - 1)}{\hat{g}(A, W, S)}$$

In this case the clever covariate simplifies to $H(A, \Delta, W) =$ a.ind, and the weights are accounted for by the "weights" option in glm.

```
offset <- qlogis(pmin(pmax(Qbar.ests[1:n],0.0005),0.9995))
eps <- coef(glm(y ~ -1 + offset(offset) + a.ind,
```

```
            weights=wgts, family=binomial))}
```

Adjust the initial $\bar{Q}_0$ estimates using the estimate $\hat{\epsilon}$

```
Q0 <- plogis(qlogis(Qbar.ests[(n+1):(2*n)]) - eps)
Q1 <- plogis(qlogis(Qbar.ests[(2*n+1):(3*n)]) + eps)
```

Estimate ATE, and $E[E[Y_1|W,S]]$ and $E[E[Y_0|W,S]]$

```
ate <- mean((Q1-Q0)*wgts)
EY1 <- mean(Q1*wgts)
EY0 <- mean(Q0*wgts)
```

Estimate influence curves for calculating the 95% CI for ATE, and the 95% CI for the treatment means $E[E[Y_1|W,S]]$ and $E[E[Y_0|W,S]]$.

```
ic.ate <- a.ind*wgts*(y-(a*Q1 + (1-a)*Q0)) + wgts*((Q1-Q0) - est)
ic1 = a*g.and.wgts*(y- Q1) + wgts*(Q1-EY1)
ic0 = (1-a)*g.and.wgts*(y- Q0) + wgts*(Q0-EY0)
ci.ate <- c(ate-qnorm(0.975)*sd(ic.ate)/sqrt(n),
    ate+qnorm(0.975)*sd(ic.ate)/sqrt(n))
ci1 <- c(EY1-qnorm(0.975)*sd(ic1)/sqrt(length(ic1)),
     EY1+qnorm(0.975)*sd(ic1)/sqrt(length(ic1)))
ci0 <- c(EY0-qnorm(0.975)*sd(ic0)/sqrt(length(ic0)),
     EY0+qnorm(0.975)*sd(ic0)/sqrt(length(ic0)))

return(list(ate=ate, ci.ate=ci.ate, EY1=EY1, EY0=EY0, ci1=ci1, ci0=ci0,
                ic.ate=ic.ate, ic1=ic1, ic0=ic0)) }
```

Using the estimates output by this function, it is straightforward to calculate the confidence intervals for vaccine efficacy using the estimates from the function and the delta method as outlined in the previous section. This is further elaborated on in the next chapter.

2.4 Role of Weights in the Statistical Formulation of Estimation of an Optimal Surrogate

In the optimal surrogate framework, the application of IPCW is straightforward and is primarily applied as the addition of appropriate weights for the various loss functions. Our proposed criterion for the optimal surrogate is again the following full-data mean squared error, but now in the IPCW-TMLE framework the second phase weighting is accounted for:

$$\psi \to MSE^{IPCW}_{P_{X,0}}(\psi) \equiv \sum_a E_{P_{X,0}} \left\{ \frac{\Delta}{\Pi_n(V)} g_0(a \mid W)(Y_a - \psi(W, a, S_a))^2 \right\}. \tag{2.3}$$

Our goal remains to minimize the weighted mean square prediction error for predicting the actual counterfactual outcome of interest, across the different treatment values, with constraint that the solution must satisfy the Prentice definition. Given a class $\mathbf{\Psi}$ of possible surrogate functions $\psi(\cdot)$ satisfying the Prentice definition by construction, the P_0-optimal surrogate in this class is then defined as

$$\psi_0^F = \arg\min_{\psi \in \mathbf{\Psi}} MSE^{IPCW}_{P_{X,0}}(\psi).$$

Once again, the choice of weight in $MSE_{P_{X,0}}$ (i.e., $g_0(a \mid W)$) along with the IPCW weight $\frac{\Delta}{\Pi_n(V)}$ does not affect the optimal solution: i.e., the optimal surrogate will be optimal for each choice of weight. The P_0-optimal surrogate ψ_0^F is given by

$$\psi_0^F(w, a, s) = E_0(Y_a \mid W = w, S_a = s),$$

which is a standard solution to a minimization problem. And, under our assumption that A

is randomized, ψ_0^F is identifiable from P_0 and can also be defined as:

$$\psi_0^F(W, A, S) = \psi_0(W, A, S) \equiv E_0(Y \mid W, A, S).$$

The application of weights to estimating $\bar{Q}_n^1$ and then the desired IPCW-TMLE would follow the steps as outlined previously in this chapter.

2.5 Application of Optimal Surrogate Methodology to Two Dengue Vaccine Efficacy Trials using Weighted SuperLearner and IPCW-TMLE

This section presents the application of IPCW-TMLE to actual dengue vaccine efficacy trial data. These results were previously presented and published in [51].

Two randomized, double-blinded, placebo-controlled, multicenter, Phase 3 trials of the identical recombinant, live, attenuated, tetravalent dengue vaccine (CYD-TDV) versus placebo were conducted in Asia [9] and Latin America [70], respectively. These trials– referred to as CYD14 and CYD15– randomized 10,275 2–14 year-old children and 20,869 9–16 year-old children, respectively, in 2:1 allocation to vaccine:placebo, with immunizations administered at months 0, 6, and 12. The primary analyses assessed vaccine efficacy (VE) against symptomatic, virologically confirmed dengue (VCD) occurring at least 28 days after the third immunization through to the Month 25 visit. Based on a proportional hazards model, estimated VE was 56.5% (95% CI 43.8–66.4) for CYD14 and 64.7% (95% CI 58.7–69.8) for CYD15.

The trials measured, from Month 13 blood samples, neutralizing antibody titers to each of the four dengue serotypes contained in the CYD-TDV vaccine using two different assays [PRNT_{50} and Microneutralization Version 2 (MNv2)]. Our analysis restricts to subjects with Month 13 titer data, which were measured in a random sample of study participants and in all subjects with the study endpoint. We use simple inverse probability weighted complete-case analysis to account for this sampling design. Each trial data set consists of

baseline covariates W (age, sex, estimated frequencies of the 4 serotypes causing dengue disease in placebo recipients in the subject's country of residence), treatment A (1=vaccine, 0=placebo), S (several variables based on the eight Month 13 titer measurements), and Y, the indicator of occurrence of the VCD endpoint between Month 13 and Month 25. The analyzed cohorts are participants observed to be free of the VCD endpoint through to the Month 13 visit with (W, A, S) measured. We treat CYD14 as the current trial and CYD15 as the future trial.

We first obtain the targeted estimated optimal surrogate $\psi_n^{\#}(W, A, S)$ for the CYD14 trial, thus obtaining TMLEs $\theta_{\psi_n^{\#}}^{TMLE,a}$ of each mean $E_0(\psi_n^{\#}(W, a, S_a))$ and of a vaccine efficacy contrast version of $\theta_{\psi_n^{\#}}^{TMLE}$, $VE_{\psi_n^{\#}}^{TMLE} = 1-\theta_{\psi_n^{\#}}^{TMLE,1}/\theta_{\psi_n^{\#}}^{TMLE,0}$ of $VE_{\psi_n^{\#}} = 1-E_0(\psi_n^{\#}(W,1,S_1))/E_0(\psi_n^{\#}(W,0,S_0))$. Second, we calculate the $\psi_n^{\#}(W^*, A^*, S^*)$ surrogate outcome values for the n^* CYD15 participants (with $\psi_n^{\#}(\cdot)$ calculated from CYD14), and, based on the CYD15 data $(W_i^*, A_i^*, S_i^*, \psi_n^{\#}(W_i^*, A_i^*, S_i^*))$, $i = 1, \cdots, n^*$, estimate the treatment-specific surrogate means $\theta_{\psi_n^{\#}}^{a}(P) = E_P\left[E_P(\psi_n^{\#}(W^*, a, S^*) \mid W^*, A^* = a)\right]$ for $a = 0, 1$ and $VE_{\psi_n^{\#}}(P) = 1 - \theta_{\psi_n^{\#}}^{1}(P)/\theta_{\psi_n^{\#}}^{0}(P)$ [estimated by $\theta_{\psi_n^{\#}}^{TMLE,a}(P) = \frac{1}{n^*}\sum_{i=1}^{n^*}\psi_n^{\#}(W_i^*, a, S_i^*)$ and $VE_{\psi_n^{\#}}^{TMLE}(P) = 1 - \theta_{\psi_n^{\#}}^{TMLE,1}(P)/\theta_{\psi_n^{\#}}^{TMLE,0}(P)$]. Lastly, because we know the Y^* values in CYD15, we compare the estimates of the surrogate parameters $\theta_{\psi_n^{\#}}^{TMLE,1}(P)$, $\theta_{\psi_n^{\#}}^{TMLE,0}(P)$, and $VE_{\psi_n^{\#}}^{TMLE}(P)$ to the TMLEs of $E_P(Y_0^*)$, $E_P(Y_1^*)$, and $\theta_P^* = VE_P^* = 1 - E_P(Y_1^*)/E_P(Y_0^*)$, respectively, calculated based on the CYD15 data (W_i^*, A_i^*, Y_i^*), to check how well the estimated optimal surrogate can be used to estimate the clinical parameters in the new setting. Wald 95% confidence intervals for the $\theta_{\psi_n^{\#}}^{a}(P)$ and $E_P(Y_a^*)$ parameters are calculated based on the influence curves, and then for $VE_{\psi_n^{\#}}(P)$ and VE_P^* using the delta method. In particular, for each $a = 0, 1$ the variance of the TMLE of $E_P(Y_a^*)$ is estimated by the sample variance of the n^* values of the efficient influence curve $\widetilde{D}^{eff,a}(\bar{Q}_{n*}^{\#1}, g_{n*})(W_i^*, A_i^*, Y_i^*)$. The delta method is applied to obtain the variance of $\log(E_P(Y_1^*)/E_P(Y_0^*))$ and hence a symmetric Wald 95% CI for this parameter, and the symmetric limits are transformed to obtain the CI for VE_P^*. The

same approach is used for $\theta^a_{\psi_n^\#}(P)$ and $VE_{\psi_n^\#}(P)$, where now the efficient influence curves are discussed in Section 2.3.

2.5.1 Targeted super-learner estimate of $\psi_0 = E_0(Y|W,A,S)$ in the CYD14 trial

We applied super-learner with 7-fold cross-validation, separately for the vaccine and placebo groups. The selection of folds was selected to be 7 instead of 10 to ensure a sufficient number of cases in each fold. Table 2.1 displays the input variables, learners, and pre-screening approaches (35 total statistical algorithms). The set of learners and screening procedures was restricted to those that allow use of inverse probability weights. Figure 2.1 shows point and 95% CI estimates of the cross-validated MSEs [62] for each individual learning algorithm as well as for discrete super-learner and super-learner.

The model fitting a logistic regression (glm) after a variable screening that disallows PRNT_{50} titers performs best (with the lowest CV-MSE) for each of the vaccine and placebo groups; these models are shown in Table 2.2. For both treatment groups the super-learner performs with similar, but slightly higher, CV-MSE. Classification accuracy is better for the vaccine than placebo group with cross-validated MSE of the super-learner 0.11 (95% CI 0.09-0.13) and 0.26 (95% CI 0.22–0.30), respectively.

Next, the TMLE $\psi_n^\#(W,A,S)$ was obtained as described in Section 1.5. To study how well this estimated optimal surrogate classifies VCD in CYD14, Figure 2.2(a) shows empirical reverse cdf plots of $\psi_n^\#(W_i, A_i = a, S_i)$ by treatment group $a \in \{0,1\}$ and VCD case-control outcome $y \in \{0,1\}$ for CYD14 data, again showing better classification in the vaccine group. Based on $\psi_n^\#(W,A,S)$, $\widehat{E}_0(Y_1) = \theta^{TMLE,1}_{\psi_n^\#} = 0.017$ (95% CI 0.016–0.019), $\widehat{E}_0(Y_0) = \theta^{TMLE,0}_{\psi_n^\#} = 0.039$ (95% CI 0.036–0.042), and $\widehat{VE}_0 = \theta^{TMLE}_{\psi_n^\#} = 55\%$ (95% CI 49–61). These estimates are close to those obtained based on (W_i, A_i, Y_i) with $\widetilde{\theta}_n^{TMLE,1} = 0.017$ (95% CI 0.014–0.021), $\widetilde{\theta}_n^{TMLE,0} = 0.039$ (95% CI 0.031–0.047), and $\widehat{VE}_0 = \widetilde{\theta}_n^{TMLE} = 55\%$ (95% CI 40–66).

Table 2.1: Input variables, screens, and learners used in the super-learner for the CYD14 dengue vaccine efficacy trial.

Input Variables	
W	Baseline demographics age (range 2–14 years), sex, empirical frequencies of the 4 serotypes in placebo group failure events by country of the participant
S	Month 13 seropositivity to each of the 4 serotypes in the CYD-TDV vaccine, and average, minimum, and maximum of the 4 titers for both PRNT_{50} and Microneutralization assays
Screens	Boldfaced courier-font screens (e.g., `screen.glmnet`) available in the SuperLearner R package available at CRAN
`screen.glmnet`	Include variables with non-zero coefficients in a standard implementation of `SL.glmnet`
screen.univar.logistic.x	Univariate logistic regression p-value < 0.10 using "x" most univariatly significant terms.
screen.corX.x	Disallow pairs of quantitative variables with $R^2 >$ "$0.x$"
screen.PRNT	Disallow Microneutralization titer variables
screen.MNv2	Disallow PRNT_{50} titer variables
Learners	Boldfaced courier-font learning algorithms (e.g., `SL.mean`) are available in the SuperLearner R package available at CRAN
`SL.mean`	$E_0(Y\|W, A = a, S)^a = \beta_a$ for $a \in \{0, 1\}$
`SL.glm`	Logistic regression with all input variables
`SL.step`	Best logistic regression model by AIC from a step-wise search
`SL.bayesglm`	Logistic regression utilizing Cauchy Bayesian priors on model parameters
`SL.polymars`	Multivariate adaptive polynomial spline regression
`Discrete SL`	van der Laan, Polley, and Hubbard (2007)
`Super Learner (SL)`	van der Laan, Polley, and Hubbard (2007)

[a] All learners were fit separately for each treatment group $A = a$ for $a \in \{0, 1\}$ as described in Section 6.1. This is explicitly stated here for `SL.mean`, and can be assumed for all other learners.

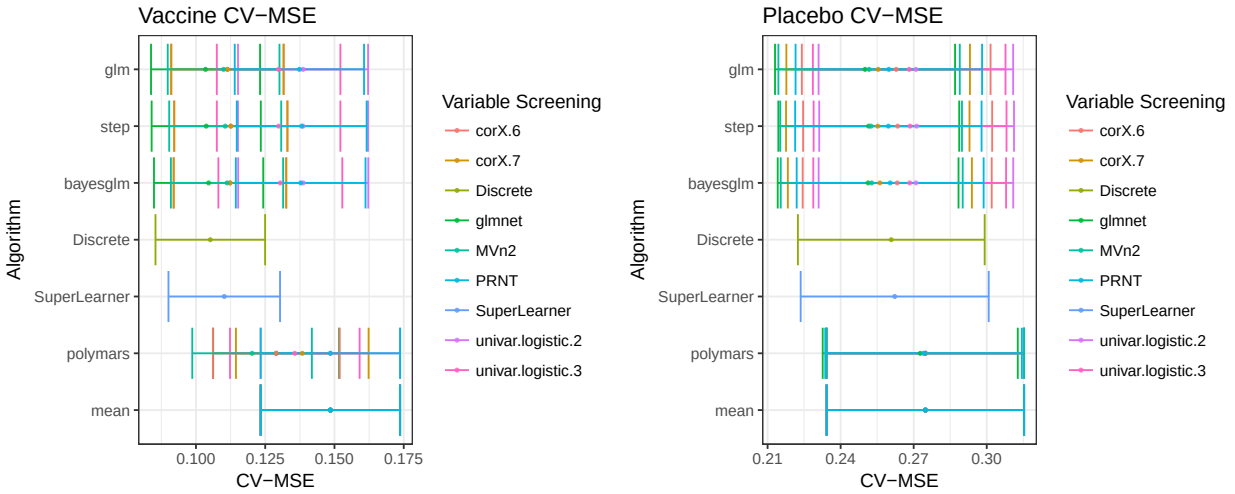

Figure 2.1: Point and 95% confidence interval estimates of cross-validated mean squared error (CV-MSE) for the vaccine and placebo groups of the CYD14 trial, all individual learners, the discrete super-learner, and the super-learner.

2.5.2 Applying the estimated optimal surrogate from CYD14 data to the CYD15 trial

Table 2.3 compares estimates of $\theta^a_{\psi_n^\#}(P)$ and of $\theta_{\psi_n^\#}(P) = VE_{\psi_n^\#}(P)$ to the estimates of $E_P(Y_0^*)$, $E_P(Y_1^*)$, and $\theta_P^* = VE_P^* = 1 - E_P(Y_1^*)/E_P(Y_0^*)$. The results show similar vaccine efficacy estimates, with $VE^{TMLE}_{\psi_n^\#}(P) = 66\%$ (95% CI 58–72) and $\widehat{VE}_P^* = \widetilde{\theta}^{TMLE}_{n*}(P) = 61\%$ (95% CI 51–69). However, the estimates of the treatment-specific surrogate means overestimate the VCD disease rates in CYD15, especially for the placebo group. The discrepancy is probably due to the imperfect adherence to the assumptions of Theorem 2, as identified through diagnostic analysis in Sections 2.5.3 and 2.5.4 .

Table 2.2: Best performing models for the vaccine and placebo groups of the CYD14 trial. For both the vaccine and placebo groups the model with the lowest CV-MSE was a logistic regression (glm) using variables selected from the screen screen.MNv2 in Table 2.1.

Model Term	Coefficient	Odds Ratio	2-Sided P-value
Vaccine Model			
(Intercept)	1.09	2.96	0.26
AGE.9.11	-0.09	0.91	0.74
AGE.12.14	-2.46	0.09	<0.01
MALE	-0.36	0.70	0.09
M13.MNv2.S1[b]	-3.62	0.03	<0.01
M13.MNv2.S2	0.77	2.16	0.02
M13.MNv2.S3	1.41	4.09	0.04
M13.MNv2.S4	-0.12	0.89	0.81
M13.MNv2.Ave[c]	3.45	31.53	<0.01
M13.MNv2.Min	-3.53	0.03	<0.01
M13.MNv2.Max	-0.59	0.55	0.28
Sero2.frequency[d]	-0.91	<0.01	<0.01
Sero3.frequency	-0.57	<0.01	<0.01
Sero4.frequency	-0.38	0.02	<0.01
Placebo Model			
(Intercept)	1.97	7.16	0.01
AGE.9.11	0.84	2.32	<0.01
AGE.12.14	-0.17	0.85	0.55
MALE	0.04	1.04	0.82
M13.MNv2.S1[b]	-1.10	0.33	<0.01
M13.MNv2.S2	0.25	1.29	0.34
M13.MNv2.S3	0.56	1.76	0.10
M13.MNv2.S4	0.06	1.06	0.84
M13.MNv2.Ave[c]	1.01	2.75	0.43
M13.MNv2.Min	-2.62	0.07	<0.01
M13.MNv2.Max	-0.25	0.78	0.51
Sero2.frequency[d]	-0.72	<0.01	<0.01
Sero3.frequency	-0.54	<0.01	<0.01
Sero4.frequency	-0.46	<0.01	<0.01

[a] The reference age category is 2–8 year olds.

[b] M13.MNv2.S1 is the binary indicator of a Month 13 positive response to serotype 1 using the MNv2 assay, with positive response defined by MNv2 serotype neutralization titer ≥ 10. M13.MNv2.S2-M13.MNv2.S4 are defined similarly.

[c] M13.MNv2.Ave, M13.MNv2.Min, and M13.MNv2.Max coefficients are per one $\log_{10}$ increase in neutralization titer value.

[d] Serotype frequency variable coefficients are per 0.10 increase in the serotype frequency of a participant's country.

Table 2.3: Comparison of inferences on the surrogate parameters in which $\theta^a_{\psi_n^\#}(P) \equiv E_P[E_P(\psi_n^\#(W^*, a, S^*) \mid W^*, A^* = a)]$ for each $a \in \{0, 1\}$ and $VE_{\psi_n^\#}(P) = 1 - \theta^1_{\psi_n^\#}(P)/\theta^0_{\psi_n^\#}(P)$ based on $(W^*, A^*, \psi_n^\#(W^*, A^*, S^*))$ versus direct inferences on the clinical dengue endpoint parameters $E_P(Y_a^*)$ and $\theta_P^* = VE_P^* = 1 - E_P(Y_1^*)/E_P(Y_0^*)$ in CYD15. Included is a summary of enrollment numbers, incidence of VCD, and number of subjects with measured titers for each study.

Surrogate Parameters Estimated by TMLEs[a]		Clinical Parameters Estimated by TMLEs[b]	
$\theta^1_{\psi_n^\#}(P)$	0.020 (95% CI 0.017–0.022)	$E_P(Y_1^*)$	0.014 (95% CI 0.012–0.017)
$\theta^0_{\psi_n^\#}(P)$	0.057 (95% CI 0.049–0.065)	$E_P(Y_0^*)$	0.037 (95% CI 0.031–0.043)
$VE_{\psi_n^\#}(P)$	66% (95% CI 58–72)	VE_P^*	61% (95% CI 51–69)

	No. Enrolled	No. VCD cases ($Y = 1$ or $Y^* = 1$)	No. with (W, A, S) or (W^*, A^*, S^*) measured[c]
Study	*Vaccine, Placebo*	*Vaccine, Placebo*	*Vaccine, Placebo*
CYD14	6851, 3424	117, 133	736, 415
CYD15	13920, 6949	184, 232	944, 587

[a] TMLEs $\theta^{TMLE,1}_{\psi_n^\#}(P)$, $\theta^{TMLE,0}_{\psi_n^\#}(P)$, and $VE^{TMLE}_{\psi_n^\#}(P) = 1 - \theta^{TMLE,1}_{\psi_n^\#}(P)/\theta^{TMLE,0}_{\psi_n^\#}(P)$.

[b] TMLEs $\widetilde{\theta}^{TMLE,1}_{n*}(P)$, $\widetilde{\theta}^{TMLE,0}_{n*}(P)$, and $\tilde{VE}_{n*}(P) = 1 - \widetilde{\theta}^{TMLE,1}_{n*}(P)/\widetilde{\theta}^{TMLE,0}_{n*}(P)$.

[c] Measured in 98.30% and 99.76% of cases with $Y = 1$ or $Y^* = 1$ for CYD14 and CYD15, respectively.

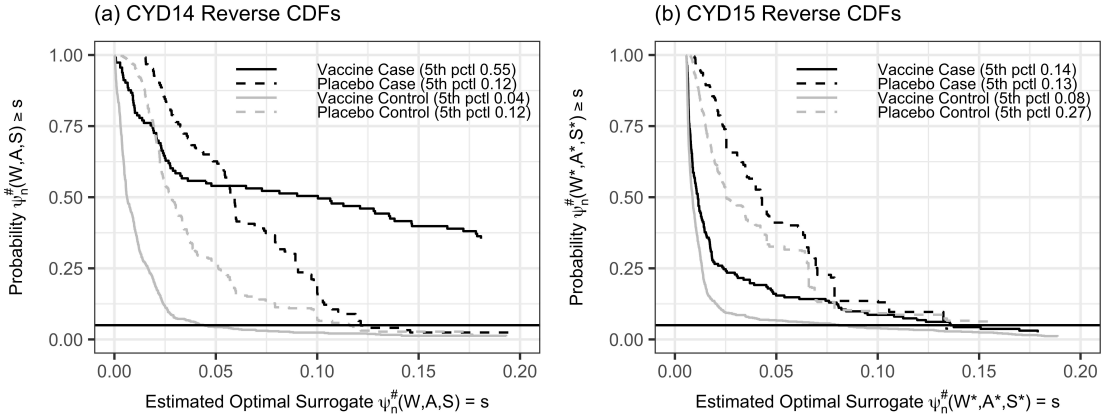

Figure 2.2: (a) Empirical reverse cumulative distribution functions (cdfs) of the estimated optimal surrogate $\psi_n^{\#}(W_i, A_i = a, S_i)$ for the CYD14 trial by vaccine/placebo assignment $A = a \in \{0, 1\}$ and dengue outcome case/control status $Y = y \in \{0, 1\}$. (b) Empirical reverse cdfs of $\psi_n^{\#}(W_i^*, A_i^* = a, S_i^*)$ for CYD15 participants by vaccine/placebo assignment $A^* = a \in \{0, 1\}$ and dengue outcome case/control status $Y^* = y \in \{0, 1\}$, where $\psi_n^{\#}(\cdot)$ was estimated from the CYD14 trial data. The results show that the surrogate better classifies dengue outcomes of participants in the original trial than in the new trial, as expected.

2.5.3 Additional analyses of the CYD14 and CYD15 dengue vaccine efficacy trial data sets

Figures 2.3–2.6 display Month 13 PRNT_{50} and Microneutralization Version 2 (MNv2) neutralization titers to the four dengue serotypes in the CYD-TDV vaccine (S) by protocol-specified age and sex covariate categories (W) and the treatment category A (where $A = 1$ for vaccine and $A = 0$ for placebo). For both CYD14 and CYD15 it is apparent that older children tend to have higher neutralization titers to all 4 serotypes than do younger children, based on both assays. Additionally, for both studies, there is an observable difference in the distributions of Month 13 neutralization titers between the vaccine and placebo groups, with higher Month 13 titers seen on average for the vaccine group. This is expected given that one of the designed purposes of vaccination is to generate neutralizing antibody responses.

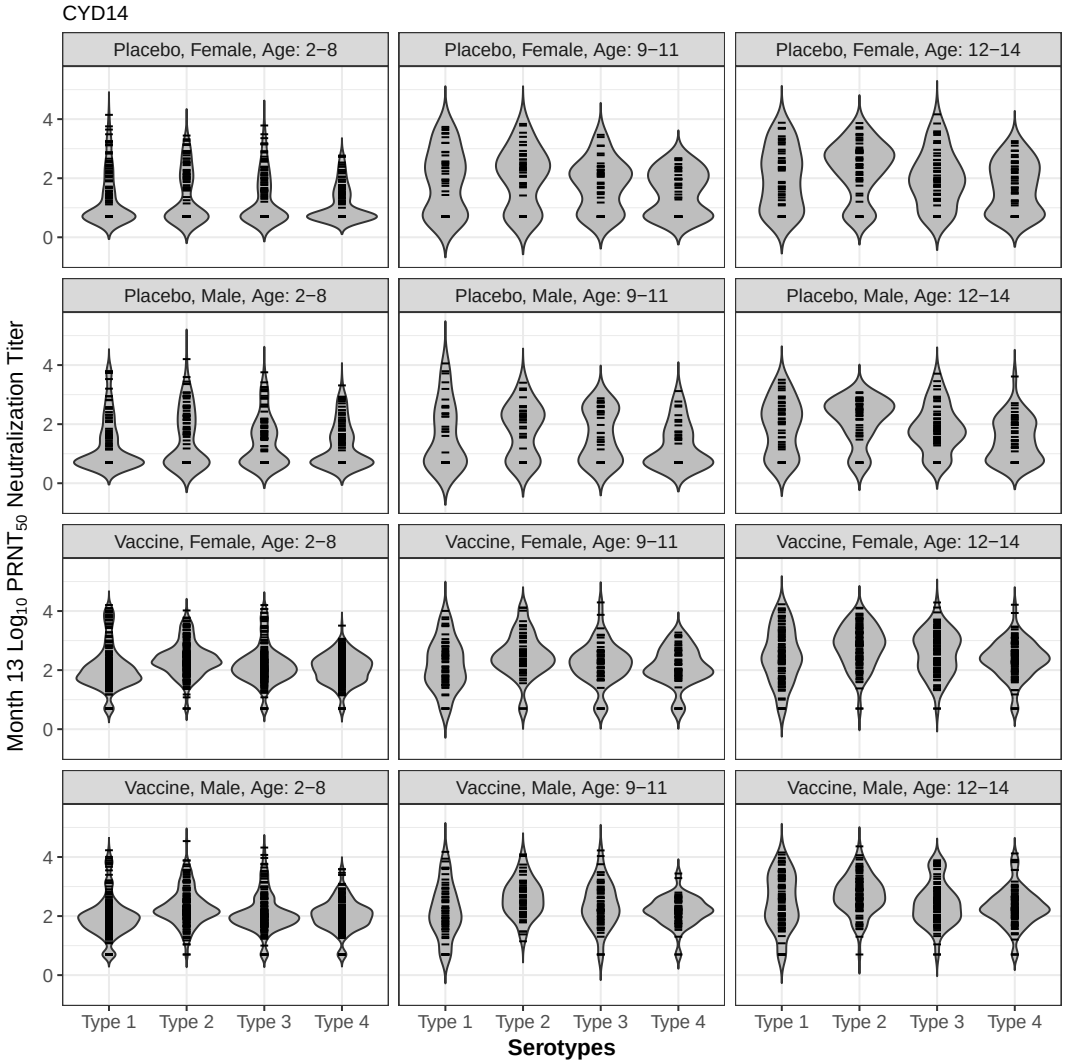

Figure 2.3: Distributions of log_{10} Month 13 neutralizing antibody titers measured by the $PRNT_{50}$ assay for each of the 4 dengue serotypes by sex and age categories for the CYD14 trial

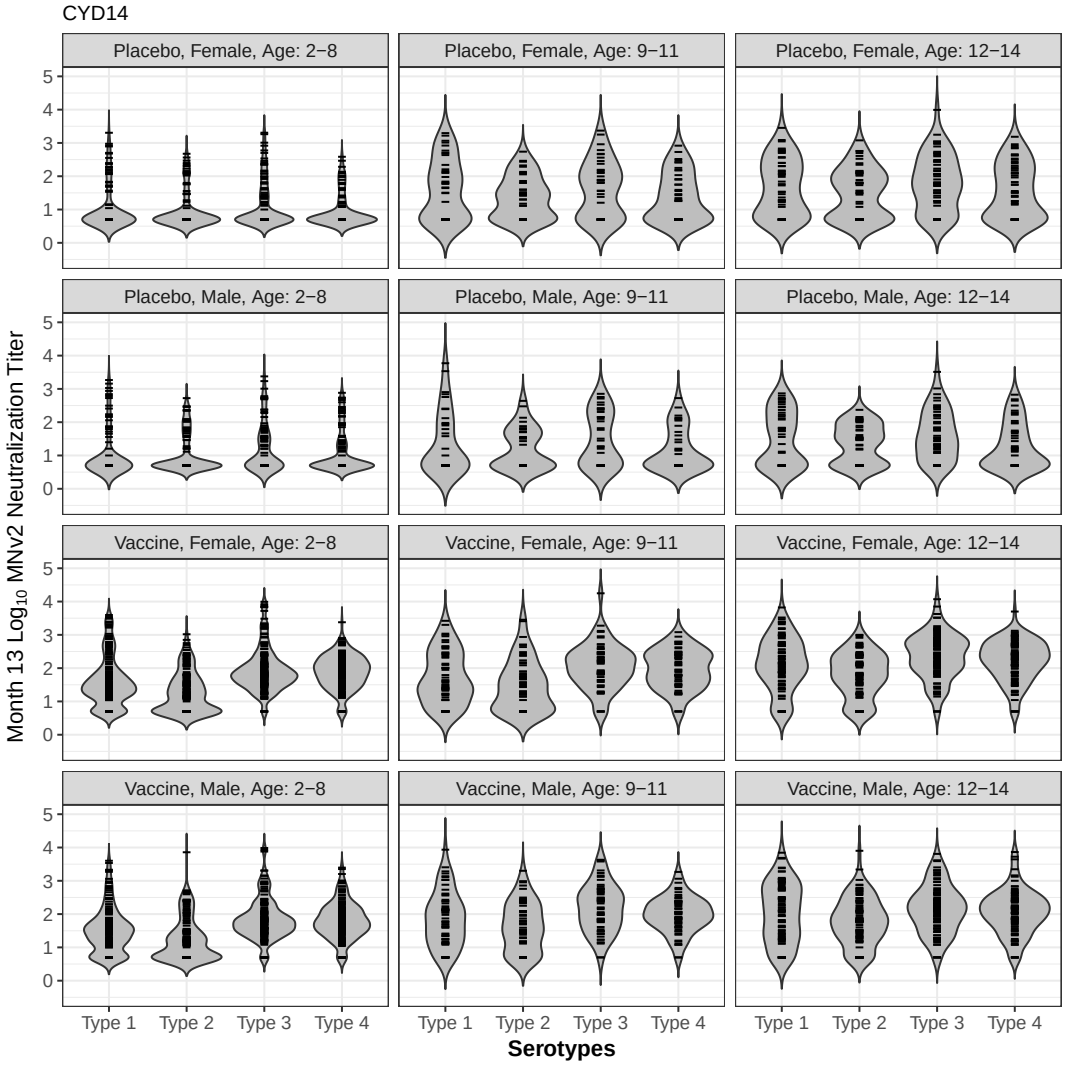

Figure 2.4: Distributions of log_{10} Month 13 neutralizing antibody titers measured by the Microneutralization Version 2 assay (MNv2) for each of the 4 dengue serotypes by sex and age categories for the CYD14 trial

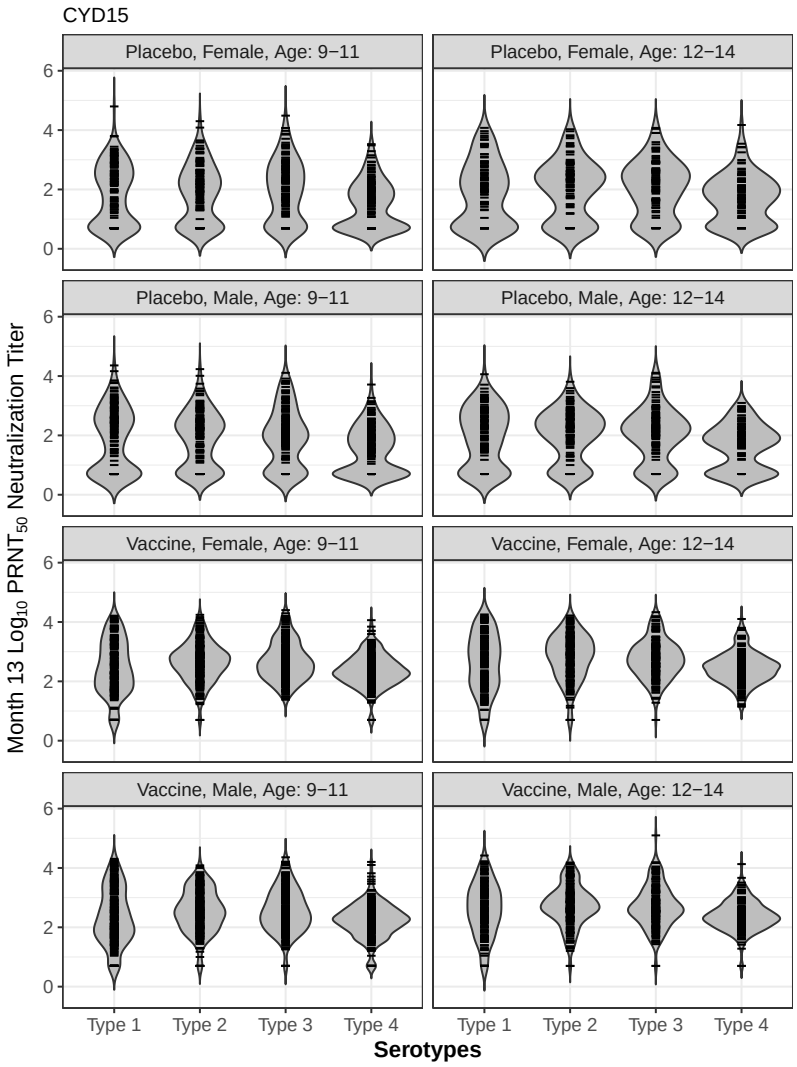

Figure 2.5: Distributions of log_{10} Month 13 neutralizing antibody titers measured by the $PRNT_{50}$ assay for each of the 4 dengue serotypes by sex and age categories for the CYD15 trial

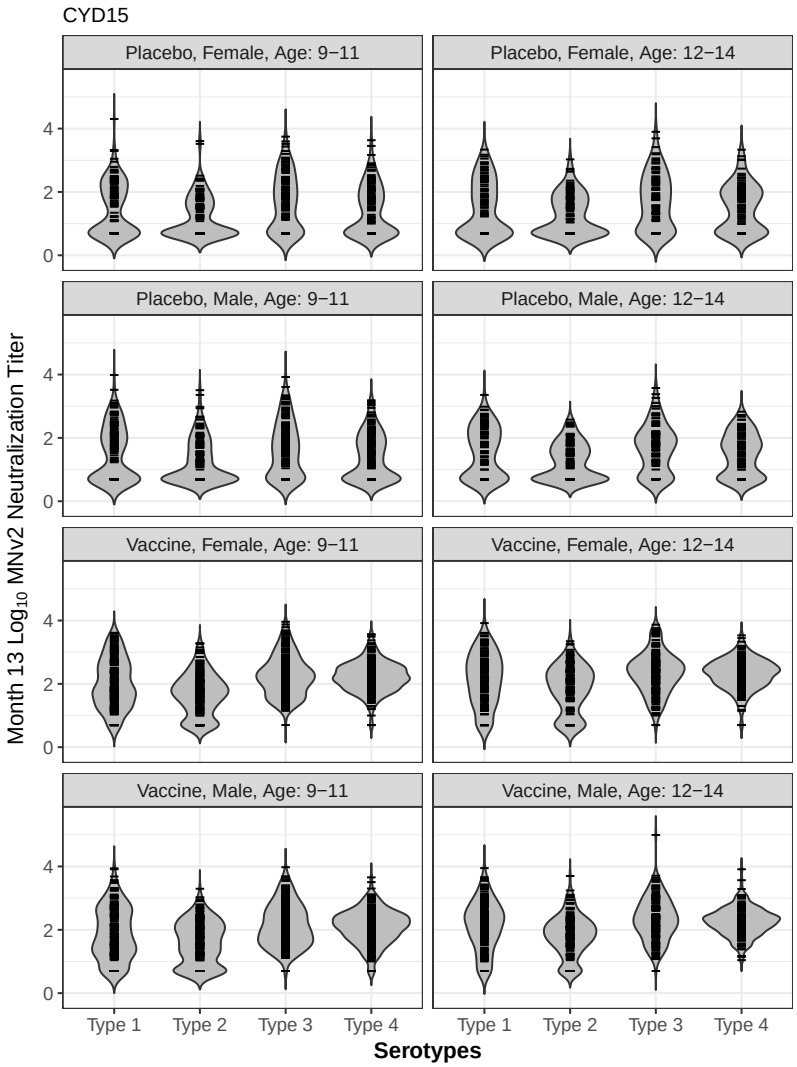

Figure 2.6: Distributions of log_{10} Month 13 neutralizing antibody titers measured by the Microneutralization Version 2 assay (MNv2) for each of the 4 dengue serotypes by sex and age categories for the CYD15 trial

2.5.4 Diagnostic checking of the conditions of Theorem 2 for the dengue vaccine efficacy trials application

In Theorem 2 we established that the estimated optimal surrogate is still a valid surrogate in a new study under the following three conditions:

1. (Randomization) A remains randomized, conditional on W,

2. (Equal Conditional Means) the conditional mean of Y given (W, A, S) is the same as the conditional mean of Y^* given (W^*, A^*, S^*), and

3. (Contained Support) the support of (W^*, A^*, S^*) is contained in the support of (W, A, S).

We check these conditions treating CYD14 as the original study and CYD15 as the new study.

Condition 1 (Randomization) is met by the fact that both CYD14 and CYD15 randomized study participants to treatment (vaccine versus placebo).

Condition 2 (Equal Conditional Means) is explored in Figures 2.7 and 2.8, which display the difference between the targeted estimated optimal surrogate $\psi_n^{\#_{14}}(W, A, S)$ built using CYD14 data and the targeted estimated optimal surrogate $\psi_{n*}^{\#_{15}}(W^*, A^*, S^*)$ built using CYD15 data. For each fixed level of the observations in CYD15, denoted by $(W^*, A^*, S^*) = (w, a, s)$, $\psi_n^{\#_{14}}(w, a, s) = \widehat{E}[Y|W = w, A = a, S = s]$ was calculated and subtracted from $\psi_{n*}^{\#_{15}}(w, a, s) = \widehat{E}[Y^*|W^* = w, A^* = a, S^* = s]$. Should the conditional mean of Y given (W, A, S) be identical to the conditional mean of Y^* given (W^*, A^*, S^*), these differences, $d(w, a, s) \equiv \widehat{E}[Y^*|W^* = w, A^* = a, S^* = s] - \widehat{E}[Y|W = w, A = a, S = s]$ should be close to zero for all observations in the CYD15 data set. As can be seen in Figures 2.7 and 2.8, these differences generally cluster around zero and are centered around zero. The values are plotted by categories of age, sex, and treatment group, and are plotted against the PRNT_{50} Month 13 (Figure 2.8) and Microneutralization Version 2 (MNv2) Month 13 neutralization

titer values (Figure 2.7). For both PRNT_{50} and MNv2 titer values, the older age category of 12–14 years for vaccinated individuals has a smaller spread of the differences $d(w, a, s)$ around 0, which was verified by the standard deviations for each category (results not shown). No other clear differences between covariate categories or between neutralization titer values are apparent.

Condition 3 (Contained Support) requires the support of (W^*, A^*, S^*) to be contained in the support of (W, A, S). The contained-support assumption holds for the age and sex covariates, because CYD14 included 2–14 year-old children and the analysis of CYD15 was restricted to 9–14 year-old children, and both studies included large numbers of male and female participants including sizable subgroups at each numeric age level. However the contained-support assumption appeared to be somewhat violated for the neutralization titer variables (Month 13 PRNT_{50} and MNv2 readouts to the four dengue serotypes). Although all titer variables had the same minimum values, and the PRNT_{50} and MNv2 serotype-specific neutralization titers were also relatively similar between the two studies, maximum titer values were slightly different in CYD15 than in CYD14. The maximum PRNT_{50} neutralization titer values for serotype 1 and for serotype 3 were 14% higher and 18% higher for CYD15 than for CYD14, respectively, but all other maximum PRNT_{50} titer values were smaller for CYD15 when compared to CYD14. For MNv2 titers, the serotype 1 maximum titer value was 9% greater, the serotype 3 maximum titer value was 17% greater, and the serotype 4 maximum titer value was 1% greater for CYD15 when compared to the maximum titers for CYD14. In sum, there are minor violations of the contained-support assumption that are expected to have a minor-to-moderate influence on the results.

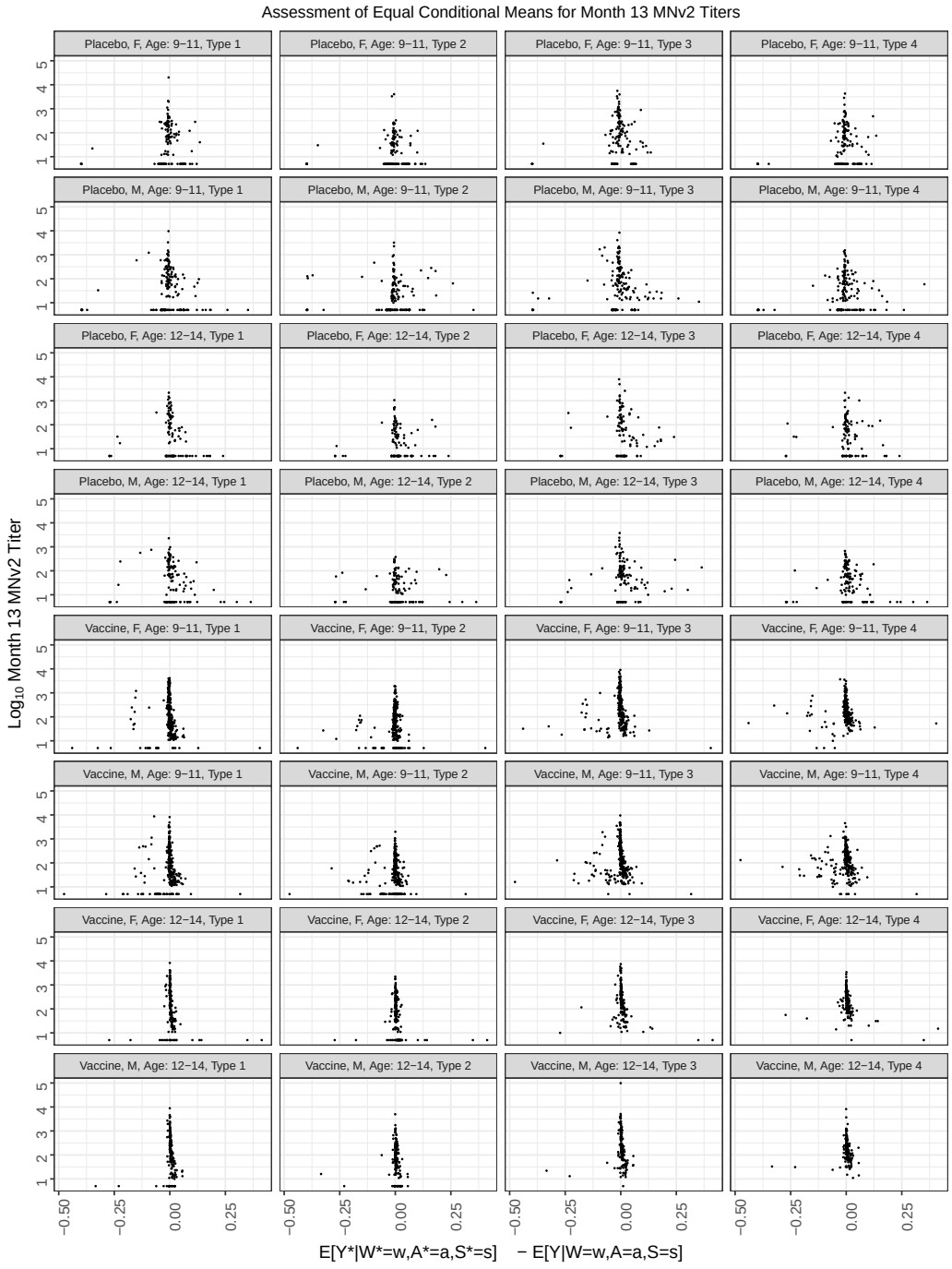

Figure 2.7: Diagnostics of the Equal Conditional Means assumption (Theorem 2): Plot of the differences (CYD15 - CYD14) in estimated optimal surrogate values for all observed values of CYD15 participants, by covariate categories and month 13 Microneutralization Version 2 titer values.

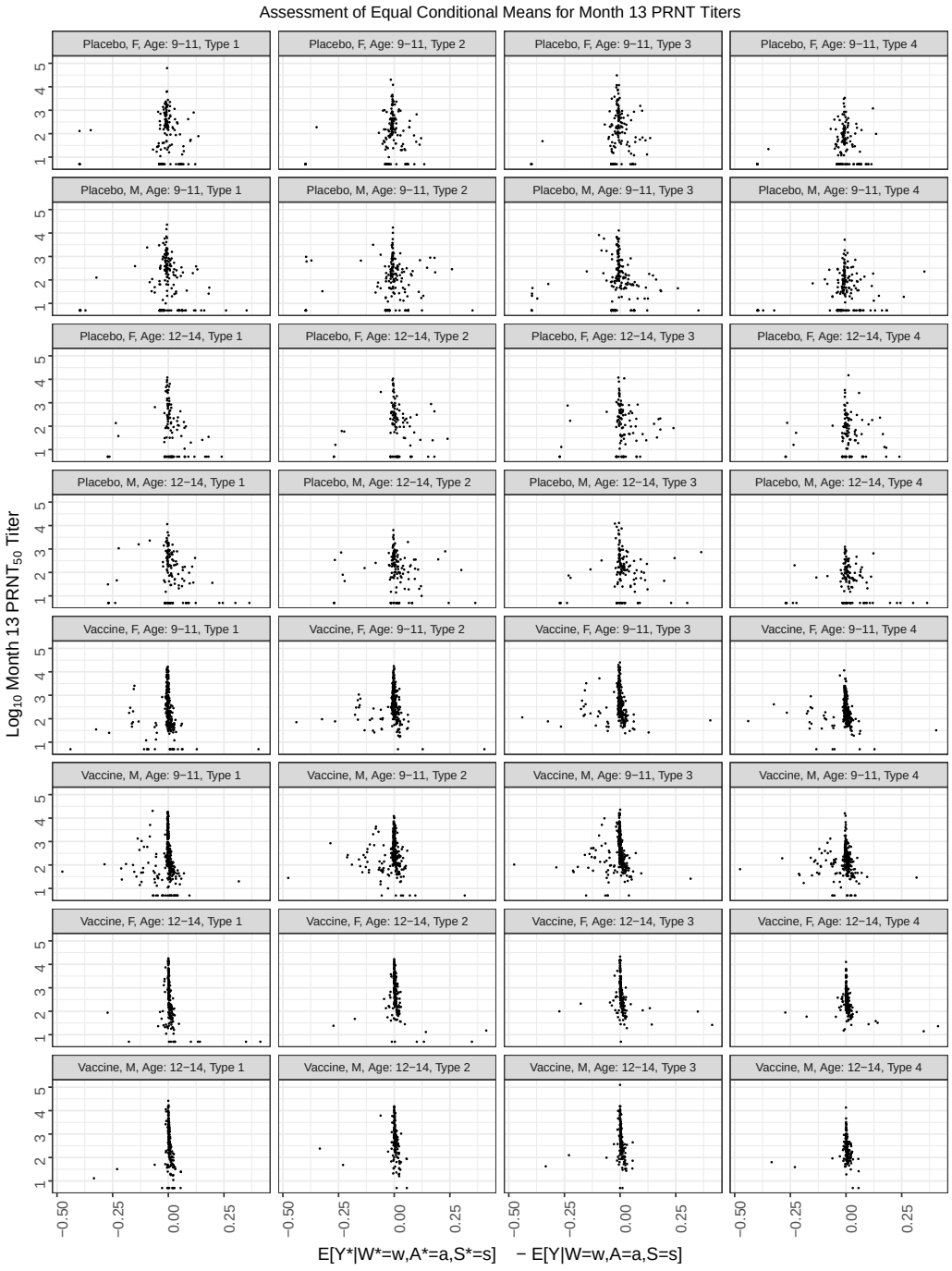

Figure 2.8: Diagnostics of the Equal Conditional Means assumption (Theorem 2): Plot of the differences (CYD15 - CYD14) in estimated optimal surrogate values for all observed values of CYD15 participants, by covariate categories and month 13 PRNT_{50} neutralization titer values.

2.6 Method for Using Algorithms in SuperLearner That Are Not Designed To Support Weights

When applying SuperLearner to two phase data or to any data requiring observation level weights, it is most desirable to have all the input learners designed to handle these types of weights. Table 2.4 presents a list of many of the available wrappers for SuperLearner and indicates which ones are not designed to handle individual weights with an asterisk. A quick perusal of this list will reveal that some desirable wrappers do not directly support individual level weights. To work around this, it has been proposed that for binary data there is a potential adjustment to the data when input into SuperLearner that will allow the researcher to use some of the wrappers that do not directly apply weights and still get valid predicted value estimates. This adjustment for basic two phase data consists of dividing the binomial outcome Y by an estimate of $\Pi_0(V)$, the probability of sampling for Phase 2. Subsequently, the SuperLearner can then be used to estimate the predicted values using an unweighted Gaussian outcome distribution, instead of weighted binomial one. This in theory would expand the number of potential algorithms that can be used to those that cannot handle weights as well as those that cannot handle binary outcomes.

Initial expectations were that this approach would work well for standard linear models, but it was uncertain if the validity of this approach would hold up for other types of algorithms, especially classification and tree-based approaches. Therefore, below we present a small simulation study to explore how this approach might work with several popular algorithms.

2.6.1 Testing the Outcome Adjustment Used to Work Around Weighting Limitations

An initial full data set of 5,000 subjects was generated with covariates W, a binary treatment indicator A, and a binary outcome Y. A single Phase I covariate W_1 is used for Phase 2 selection and is generated with a binomial distribution, where $W_1 \sim Binomial(p = 0.5)$. Phase

2 covariates, $W_2 - W_5$ were generated as continuous covariates with $W_k \sim Normal(0, 1)$, $k = 2, \ldots, 5$). The binary treatment indicator A was generated as $A \sim Binomial(p = 0.67)$, and the binary outcome was generated as $Y_i \sim Binomial(p = p_i)$ where p_i was calculated via

$$p_i^{0.25} = expit\left(-2A_i + 2W_{1_i} - W_{2_i} - 2W_{3_i}^2 + 4W_{2_i}^3\right) \tag{2.4}$$

for the simulations in which outcome rate was 25% and

$$\begin{aligned} p_i^{0.02} = expit[&-6A_i - 3W_{1_i} - 6W_{2_i}^2 - 6sin(W_{2_i})^2) - 4Asin(W_{2_i}^2) \\ &+ 8log(W_{3_i}^2) - 2W_{4_i} - 2log(W_{3_i}^2(1 - W_{4_i}))] \end{aligned} \tag{2.5}$$

for the simulations in which outcome rate was 2%.

For both data generating distributions (disease rate 25% and 2.5%), 20 training and test data sets were generated and the predicted disease rate generated by each individual learner was estimated using an array of algorithms in SuperLearner. For each data set a Phase 2 sample was drawn that included 40 cases and 400 controls. Learners tested for performance with this adjustment included some that are designed to handle observation-level weights and some that are not. Those tested included SL.mean, SL.step, SL.bayesglm, SL.glm, SL.gam, SL.gam.3, SL.gam.4, SL.glm.interaction, SL.glmnet, SL.lm, SL.nnet, SL.randomForest, SL.rpart, SL.svm, SL.polymars, SL.gbm, and SL.nnls. The SuperLearner was used to estimate the predicted disease rate in thee ways: 1) for a phase 2 sample of 440 subjects (40 case, 400 controls) using a weight adjusted binary outcome of $Y/\Pi_n(V)$, 2) for the same phase 2 sample of 440 subjects using the binary outcome Y and observation-level weights $(1/\Pi_n(V))$, and 3) for the entire, unweighted data set using Y as the outcome. For 1) the SuperLearner treated the adjusted outcome $Y/\Pi_n(V)$ as Guassian, while for 2) and 3) the SuperLearner treated the outcome Y as Binomial. For 1), an additional adjustment was

made post-SuperLearner in order to transform the subsequent predicted values back to the original scale. Results for the predicted probability of disease are presented for the training data set and for a n=10,000 validation data set in Tables 2.4 and 2.5.

Preliminary results indicate that this approach may be effective for some learners, but maybe not for others. This method does appear to allow a larger array of learners because some that are not designed to handle binary outcomes can now be used for the adjusted (and assumed Gaussian) data. In Tables 2.4 and 2.5 we see the estimated means and standard errors for disease rate as estimated by each individual algorithm. Those based on a linear model appear to do well.

A closer inspection of the methodologies behind the algorithms may explain why some algorithms benefit from this approach and other may not. It is not surprising that the nnls (non-negative least squares) algorithm did not work well because our data generating distribution by definition has negative coefficients. Support vector machines (svm), using the default algorithm, appears to struggle with the smaller sample sizes of Phase 2 data and doesn't do well at all with the $Y/\Pi_n(V)$ outcome. Also of note is that neural nets (nnet) and RandomForest (randomForest) both perform better using the adjusted outcome $Y/\Pi_n(V)$ than when using a binary outcome (both weighted Phase 2 and un-weighted complete data). This was not necessarily expected because Random Forests should be able to handle binomial data and neural networks can handle a response that is a factor. Further investigation into the particular algorithms might shed light on these results more. It is comforting to note that the average estimated rates built on the training set match the estimates generated on the independent and larger validation data sets.

Table 2.4: Training & Validation Data Mean (SE) Predicted Values; Rate 25%. Numbers in bold are more than 10% different than the true rate (0.253 (0.012)).

	Adjusted Outcome	Weighted Algorithm	Complete Data
Training Data Set			
SL.mean	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.step	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.bayesglm	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.glm	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.gam	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.gam.3	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.gam.4	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.glm.interaction	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.glmnet	0.253 (0.012)	0.253 (0.012)	0.253 (0.012)
SL.lm	0.253 (0.012)	0.276 (0.012)	0.275 (0.011)
SL.nnet*	0.253 (0.012)	**0.080** (0.113)	**0.159** (0.122)
SL.randomForest*	0.255 (0.012)	**0.213** (0.009)	0.239 (0.012)
SL.svm*	**0.049** (0.014)	**0.168** (0.017)	0.253 (0.011)
SL.nnls*	**0.059** (0.005)	**0.301** (0.023)	**0.307** (0.013)
Validation Data Set			
SL.mean	0.253 (0.014)	0.253 (0.012)	0.253 (0.012)
SL.step	0.254 (0.015)	0.253 (0.016)	0.255 (0.011)
SL.bayesglm	0.254 (0.015)	0.253 (0.016)	0.255 (0.011)
SL.glm	0.254 (0.015)	0.253 (0.016)	0.255 (0.011)
SL.gam	0.257 (0.018)	0.251 (0.016)	0.255 (0.011)
SL.gam.3	0.258 (0.02)	0.247 (0.017)	0.255 (0.011)
SL.gam.4	0.259 (0.021)	0.242 (0.017)	0.255 (0.01)
SL.glm.interaction	0.251 (0.019)	0.243 (0.017)	0.255 (0.011)
SL.glmnet	0.254 (0.015)	0.253 (0.015)	0.255 (0.011)
SL.lm	0.254 (0.015)	0.276 (0.011)	0.277 (0.01)
SL.nnet*	0.252 (0.051)	**0.071** (0.101)	**0.162** (0.124)
SL.randomForest*	0.263 (0.023)	**0.155** (0.009)	0.242 (0.011)
SL.svm*	**0.047** (0.012)	**0.152** (0.01)	0.256 (0.011)
SL.nnls*	**0.059** (0.006)	**0.301** (0.017)	**0.308** (0.01)

* Not designed to handle observation-level weights, therefore weights not properly applied in the "Weighted Algorithm" approach.

Table 2.5: Training & Validation Data Mean (SE) Predicted Values; rate 2.5%. Numbers in bold are more than 10% different than the true rate (0.025 (0.003)).

	Adjusted Outcome	Weighted Algorithm	Complete Data
Training Data Set			
SL.mean	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.step	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.bayesglm	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.glm	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.gam	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.gam.3	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.gam.4	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.glm.interaction	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.glmnet	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)
SL.lm	0.025 (0.003)	0.026 (0.003)	0.026 (0.003)
SL.nnet*	0.025 (0.003)	**0.000** (0.000)	**0.000** (0.000)
SL.randomForest*	0.024 (0.003)	**0.036** (0.003)	0.022 (0.002)
SL.svm*	**0.008** (0.002)	**0.046** (0.006)	0.024 (0.003)
SL.nnls*	**0.035** (0.007)	**0.388** (0.088)	**0.449** (0.047)
Validation Data Set			
SL.mean	0.025 (0.002)	0.025 (0.003)	0.025 (0.003)
SL.step	0.025 (0.002)	0.025 (0.003)	0.025 (0.003)
SL.bayesglm	0.025 (0.002)	0.025 (0.003)	0.025 (0.003)
SL.glm	0.025 (0.002)	0.025 (0.003)	0.025 (0.003)
SL.gam	0.025 (0.002)	0.026 (0.003)	0.025 (0.002)
SL.gam.3	0.025 (0.003)	0.027 (0.005)	0.024 (0.002)
SL.gam.4	0.025 (0.003)	0.029 (0.006)	0.024 (0.002)
SL.glm.interaction	0.025 (0.002)	0.029 (0.004)	0.025 (0.002)
SL.glmnet	0.025 (0.002)	0.025 (0.003)	0.025 (0.003)
SL.lm	0.025 (0.002)	0.027 (0.003)	0.026 (0.003)
SL.nnet*	0.026 (0.004)	**0.000** (0.000)	**0.000** (0.000)
SL.randomForest*	0.024 (0.002)	**0.055** (0.005)	0.021 (0.001)
SL.svm*	**0.007** (0.002)	**0.059** (0.007)	0.025 (0.002)
SL.nnls*	**0.034** (0.008)	**0.379** (0.096)	**0.445** (0.048)

* Not designed to handle observation-level weights, therefore weights not properly applied in the "Weighted Algorithm" approach.

Chapter 3

SIMULATION STUDY COMPARING IPCW-TMLE METHODOLOGY TO A STANDARD METHOD OF TWO-PHASE DATA ANALYSIS

3.1 Introduction

Through a simulation study, this chapter compares the finite-sample performance of inverse probability of censoring weighted targeted minimum loss estimation (IPCW-TMLE), as proposed by Rose and van der Laan [55], to a standard method of two-phase sample data analysis, Breslow-Holubkov (BH) constrained maximum likelihood logistic regression [5]. IPCW-TMLE methodology provides doubly robust estimation of statistical or causal parameters measuring the association or effect of an expensive exposure variable on the outcome by using TMLE [55] and estimated inverse probability weights to account for the two-phase sampling of the exposure variable. In their paper, Rose and van der Laan present simulation studies exploring the performance of IPCW-TMLE on phase 2 samples compared to TMLE on the full cohort. Considering only the risk difference, they explored various sizes of phase 2 samples and proportions of cases to controls; however they did not compare their results to other existing methods of analysis for two-phase sampled data. Additional evaluation of relative performance of IPCW-TMLE versus established methods for two-phase study designs is needed.

IPCW-TMLE is potentially advantageous due to its double robustness, meaning that consistent estimation of parameters of interest (marginalized exposure-specific means and contrasts in these means) is achieved if either the two-phase sampling probability model, or the outcome regression linking outcome to the exposure variable of interest and other

covariates, is correctly specified. This type of robustness is appealing for two-phase sampling studies, because typically the sampling probability model can be correctly specified, due to investigator-control of the sampling design, whereas it may be very difficult to correctly specify the outcome regression model (and to verify correct specification), especially when adjusting for several covariates including continuous covariates. IPCW-TMLE fits this data set-up in providing consistent estimation allowing the outcome regression to be misspecified. In contrast, the BH approach requires a correctly specified outcome regression model in order to provide consistent estimation of the same parameters of interest. Moreover, use of nonparametric super-learning for the outcome regression as part of IPCW-TMLE may provide superior estimation of this regression, which in theory can provide improvements in both bias and efficiency. In applications where only a single discrete categorical covariate is studied, BH would be nonparametric; however, in many applications it is desirable to control for potential confounding factors including continuous covariates. In these settings, BH would require a correctly specified parametric model (logistic regression) including all of the independent variables (potential confounders and the exposure variable of interest) in order to obtain consistent estimation. In our simulation study, we adjust for continuous variables and investigate misspecification of the logistic regression model.

Our simulation study seeks to address the following questions in two-phase sampling studies: (1) How does performance of the IPCW-TMLE and BH methods for estimation and inference of marginalized exposure-specific means compare as quantified by bias, 95% confidence interval coverage probability, Monte Carlo standard error, and ratio of root mean squared error? (2) How do these methods compare in classification accuracy of the misspecified outcome regression models fit by each method, as quantified by validation data set area under the ROC curve (AUC)? By exploring these operating characteristics, this simulation study seeks to determine if IPCW-TMLE methodology increases robustness, efficiency, and power over the standard BH method under the conditions studied. In addition, both meth-

ods are applied to three vaccine efficacy clinical trial data sets for assessing biomarkers as correlates of infection or disease.

3.2 Methods

3.2.1 Data Structure of Two-Phase Sampling Studies

The two-phase sampling design applies a case-control approach to an existing cohort or data set in order to create a sample of the data set from which to obtain estimates of a parameter of interest that would reflect estimates of the same parameter obtained on the full data set. It is always desirable to obtain estimates from the full data set when possible; however, the two-phase sampling approach allows resource-efficient analysis when factors, such as cost and accessibility to measurements, prohibit obtaining complete measurements on the full data set. This design often allows for statistically efficient estimation of parameters of interest with considerable savings in cost and time [75]. The large literature on two-phase sampling data analysis includes Neyman [43], Mantel [39], Breslow, Lubin, and Marek [3], Breslow and Cain [4], Flanders and Greenland [15], Ernster [14], Chatterjee, Chen, and Breslow [11], Hak [26], Vittinghoff and Bauer [71], Wang et al. [72], Breslow et al. [6], Fong [16], and Yang et al.[76], several of which provide alternatives to IPCW-TMLE for endeavoring to increase robustness and/or efficiency of analyses by leveraging information in phase one data from participants not sampled into phase two.

At phase 1 the researcher begins with an independent sample of N_1 cases (i.e., with $Y = 1$) and N_0 controls (i.e., with $Y = 0$) from (infinite) sub-populations of interest and obtains values of phase 1 variables, which include a discrete categorical variable V with J levels that is used for selection of participants for measurement of phase 2 variables, other variables X_1 that may be of interest to adjust for in modeling, and the binary outcome of interest, Y. In general, phase 1 variables are assumed to be measured and available for all participants, and hence we assume no missing values for V, X_1, and Y. Then, two-phase

sampling is defined by random sampling of participants within each cell of the 2 x J matrix that cross-classifies the covariate levels of V with outcome status Y. In a scenario in which there is no stratification variable V, the structure simplifies to a case-control study. This two-phase sampling design is further described in Breslow et al. [6]. We denote the full set of phase 1 input variables as $X = (V, X_1)$.

We assume the phase 1 selection variable V has possible values $v_1, \ldots, v_J$. At phase 1, the number of sampled individuals with $Y = i$ and $V = v_j$ is represented as N_{ij}. For phase 2, for each $(i, j) \in \{0, 1\} \times \{1, \ldots, J\}$ the researcher randomly samples n_{ij} individuals from among the N_{ij} individuals, from whom the phase 2 exposure of interest A and the remaining phase 2 covariates W are measured, which can either be discrete or continuous. We use Δ, with realization δ, to indicate selection for the second phase and therefore present this sampling as a missing data structure with $\delta = 1$ indicating non-missing values. At the second phase, the researcher now has a set of potential modeling variables that can include all phase 1 variables $X = (V, X_1)$, additional phase 2 variables W, and the phase 2 exposure of interest A. The observed data vector for each participant can be represented as $O = (V, X_1, Y, \Delta, \Delta A, \Delta W)$ with $N = N_1 + N_0$ iid copies of O, distributed as P, the probability distribution of O. Estimation and inference objectives can then be based on the $n = \sum_{i=0}^{1} \sum_{j=1}^{J} n_{ij}$ values of O from the phase 2 data set consisting of all participants with $\delta = 1$. For application of the BH estimator, cell counts from phase 1 (the N_{ij}) are used to appropriately adjust estimates for the two-phase sampling, whereas for the IPCW-TMLE estimator, IPCW weights are used for appropriate adjustment (Section 3.2.4). The random sampling from each cell of the $2 \times J$ matrix can be done without replacement or via Bernoulli sampling; we focus on Bernoulli sampling throughout this chapter. Throughout we assume the two-phase sampling is missing at random, meaning that the probability of phase-two sampling, $\pi(V, Y) = P(\Delta = 1|V, Y)$, does not depend on any other participant variables (observed or unobserved) other than V and Y.

3.2.2 Parameters of Interest

We first define two statistical parameters of interest from which a variety of contrasts can be constructed to answer questions of interest. Let $\Psi_a = E[E[Y|A = a, W, X]]$ for each $a = 0, 1$, where E is expectation under the distribution P and A is the binary exposure indicator; we refer to Ψ_a as the marginalized exposure-specific mean. Then, for any continuous and differentiable contrast function $h(x, y)$ satisfying $h(x, y) = 0$ if and only if $x = y$ and $h(x, y) < 0$ for $x < y$, one can estimate a contrast parameter $\theta = h(\Psi_1, \Psi_0)$ by $\hat{\theta} = h(\hat{\Psi}_1, \hat{\Psi}_0)$, where $\hat{\Psi}_a$ is an estimate of Ψ_a. Three common quantities of interest using this approach include $h(x, y) = x - y$, the marginalized risk difference (RD); $h(x, y) = x/y$, the relative marginalized risk (RR); and $h(x, y) = [x/(1 - x)] \,/\, [y/(1 - y)]$, the marginalized odds ratio (OR). Furthermore, using estimates of $var(\Psi_1)$ and $var(\Psi_0)$, application of the delta method yields the variance of $\hat{\theta}$ for any contrast function $h(x, y)$. For example, using the delta method it is straightforward to estimate confidence intervals for RR, estimating $var(\widehat{log(RR)}) = (1/\hat{\Psi}_1)^2 var(\hat{\Psi}_1) + (1/\hat{\Psi}_0)^2 var(\hat{\Psi}_0)$ from estimates $\hat{\Psi}_1$, $\hat{\Psi}_0$, $var(\hat{\Psi}_1)$ and $var(\hat{\Psi}_0)$, and then back-transforming to the RR scale.

The parameters Ψ_a and θ are statistical parameters, not requiring causal assumptions for identifiability and estimation. However, causal parameter are more interpretable for certain objectives, and under causal assumptions $\Psi_a = E(Y_a)$, where $E(Y_a)$ is a causal parameter – the exposure-a-specific mean – with interpretation as the probability of $Y = 1$ in a hypothetical scenario in which the whole study population were assigned exposure $A = a$. The (standard) causal assumptions are: (i) SUTVA, (ii) Consistency, (iii) Ignorability (i.e., A is randomized conditional on X and W), and (iv) Positivity (i.e. $P(A = 1|X = x, W = w) > 0$ and $P(A = 0|X = x, W = w) > 0$ for each possible value of (x, w)). Under these causal assumptions, the marginalized RD is also the average exposure causal effect (i.e., the commonly studied ATE), and the marginalized RR and OR parameters are the causal relative risk and causal odds ratio, respectively.

Another potential advantage of IPCW-TMLE is improved efficiency; Rose and van der Laan (2011) showed that for a discrete phase one variable V and a vector of phase-two variables W, if a nonparametric estimator of $\pi(V, Y)$ is used, and the estimator for the outcome regression is consistent, then, under regularity conditions, IPCW-TMLE provides an asymptotically efficient estimator of each Ψ_a (and contrasts θ) accounting for the iid copies of $O = (V, Y, \Delta, \Delta A, \Delta W)$. Therefore, another potential advantage of IPCW-TMLE is improved efficiency, and our simulations and applications use a discrete phase one variable and a nonparametric estimator of $\pi(V, Y)$.

3.2.3 Breslow-Holubkov (BH) Approach

Breslow and Holubkov [5] developed a semiparametric maximum likelihood estimation approach for two-phase studies using a logistic regression model. With a binary exposure A, this model specifies a linear logistic regression model $P(Y = 1|A, X, W) = p_1(A, X, W; \beta)$, for example

$$p_1(A, X, W; \beta) = (1 + exp[-(A\beta_A + X^T\beta_X + W^T\beta_W)])^{-1},$$

which could be enriched with additional variables such as interaction terms. The BH approach assumes this logistic regression holds for all values of (a, x, w), and leaves the distributions of X and W completely unspecified. BH fits this model by constrained maximum likelihood estimation, which requires the marginal probabilities of being a case $(Y = 1)$ or a control $(Y = 0)$ at phase 1, and of being a case or control at phase 2 in each stratum j, to be fixed by design. Estimates of the model parameters using the BH approach can be obtained using the R package *osDesign* [28].

Predicted probabilities of outcome can be derived from the estimated model and then used to estimate the parameters of interest $\Psi_a = E[E(Y|A = a, W, X)]$. An estimate of Ψ_a

for $a = 0, 1$ by the BH approach can be obtained as

$$\hat{\Psi}_a = \frac{1}{N_a} \sum_{A_k=a} \frac{1}{\pi_n(V_k, Y_k)} \widehat{E}[Y_k | A_k = a, W_k, X_k], \tag{3.1}$$

where $\widehat{E}[Y_k | A_k = a, W_k, X_k]$ is the fitted value from the logistic regression model, and the sum is over individuals in group a that have phase-two data. An application of the multivariate delta method yields the variance of $\hat{\Psi}_a$ for each $a = 0, 1$.

In equation (3.1), $\pi_n(V, Y)$ is an estimator of the phase-two sampling probabilities $\pi(V, Y) = P(\Delta = 1 | V, Y)$, which may be nonparametric (i.e., empirical fractions with numerators based on phase-two individuals and denominators based on phase-one individuals) or fitted values from a parametric model such as logistic regression. These probabilities are often known because the researcher defines sampling criteria based on V and Y. However, superior performance may typically be obtained by using an estimate $\pi_n(V, Y)$ of $\pi(V, Y)$; moreover there is often also happenstance missingness of phase-two data, further motivating use of estimates.

3.2.4 IPCW-TMLE Approach

The first step in the application of IPCW-TMLE estimates $\bar{Q}_0 = E[Y | A, W, X]$, with estimator denoted by $\bar{Q}_n = E_n[Y | A, W, X]$. TMLE generally uses a SuperLearner estimator, a nonparametric, loss-based, ensemble, and cross-validation based estimator [63, 48, 47]. SuperLearner has an oracle property that, for a pre-specified loss function, it should perform as well as the unknown best individual estimator in the specified ensemble of estimators [64, 65]. Using SuperLearner to estimate $\bar{Q}_0$ requires careful attention to the two-phase sampling probabilities. Our IPCW-TMLE approach applies SuperLearner with each individual estimator of $\bar{Q}_0$ employing observation-level weights equal to reciprocals of estimated probabilities of sampling individuals into phase two. Additionally, the K-fold cross-validation

scheme is designed to ensure that similar numbers of cases are included in each fold, particularly to avoid folds with no events that can readily occur in applications with rare events. The library of estimators includes a wide variety of ways to input (A, W, X) (e.g., interactions, transformations) as potential predictors of Y. The Superlearner estimator $\bar{Q}_n$ of $\bar{Q}_0$ is defined relative to a pre-specified loss function, as the optimal convex combination of these estimators that minimizes the cross-validated risk.

Once the initial estimate $\bar{Q}_n$ is obtained, in order to "target" the estimator of Ψ_a to reduce bias and obtain a valid estimate of the standard error, the exposure/treatment assignment mechanism $g_0(a|W, X) = P(A = a|W, X)$ for $a = 0, 1$ is also estimated. In a randomized clinical trial, g_0 is typically known by study design, but can also be estimated using SuperLearner, which tends to improve performance by accounting for finite-sample random imbalances in confounders [64]. Our applications have a two-phase exposure variable A that is not randomized, in which case it is important to estimate g_0. Therefore, SuperLearner is used to estimate g_0 by setting exposure A as the outcome modeled using algorithms that are functions of the covariates W and X.

For IPCW-TMLE, the phase-two sampling probabilities $\pi(V, Y) = P(\Delta = 1|V, Y)$ can be estimated in the same way as for the BH method (empirical fractions or parametrically), or using Superlearner. The IPCW weight for individual k in the phase-two sample is $1/\pi_n(V_k, Y_k)$. In our applications, we estimate the $\pi_n(V_k, Y_k)$ empirically from the frequencies of sampling within each stratum.

For estimation of each Ψ_a for $a = 0, 1$, the weights $1/\pi_n(V_k, Y_k)$ are used in the Superlearner estimator $Q_n(A = a, W, X)$, as well as in the estimator $g_n(1|W, X)$. At the bias-reduction targeting step, the same weights are also used in the selected loss function, the linear logistic working submodel, the final IPCW-TMLE plug-in estimator [shown in equation (3.2) below], and the estimation of standard errors using influence curves as described in [55]. A working submodel is selected to calculate the amount of fluctuation of $\bar{Q}_n$

to iteratively adjust it to obtain an efficient estimate of the parameter of interest Ψ_a. This step uses a weighted logistic regression with outcome Y, a univariate covariate $H_n^*(A, W, X)$ specific to the target parameter (sometimes termed the "clever covariate"), and offset constants $K(W, X)$ that are held fixed (i.e., as intercept) in the weighted maximum likelihood linear regression fit. To estimate $\bar{Q}^1 = E[Y|A = 1, W, X]$ and $\bar{Q}^0 = E[Y|A = 0, W, X]$ separately, individual clever covariates and offset constants specific to each of the two parameters are used. Specifically, for estimating $\bar{Q}^1$, a weighted logistic regression is fit to the subset of data with $A = 1$ using the model $Y = K_1(W, X) + \epsilon_1 * H_n^*(A, W, X)$ where $K_1(W, X) = logit(\bar{Q}_n(A = 1, W, X))$, $H_n^*(A, W, X) = A/g_n(A|W, X)$, and the weights are $1/\pi_n(V_k, Y_k)$. The obtained estimate of ϵ_1, ϵ_{n1}, is saved and used to update $\bar{Q}_n^1$ to the value $\bar{Q}_n^{1*} = expit\left[logit(\bar{Q}_n(1, W, X)) + \epsilon_{n1}/g_n(1|W, X)\right]$. This process is iterated using the new estimate $\bar{Q}_n^{1*}$ in place of $\bar{Q}_n^1$ until convergence, which, for this particular parameter, is achieved after one iteration. At convergence the final $\bar{Q}_n^{1*}$ equals the IPCW-TMLE of $\bar{Q}^1$.

A similar approach is used for estimating $\bar{Q}^0 = E[Y|A = 0, W, X]$. For the subset of data with $A = 0$ we fit the logistic regression model $Y = K_0(W, X) + \epsilon_0 * H_n^*(A, W, X)$ with $H_n^*(A, W, X) = (1 - A)/(1 - g_n(1|W, X))$ and offset constants $K_0(W, X) = logit(\bar{Q}_n(A = 0, W, X))$, once again with weights $1/\pi_n(V_k, Y_k)$. This process is iterated until convergence, from which the final estimate $\bar{Q}_n^{0*}$ of $\bar{Q}^0$ is obtained.

The final IPCW-TMLE estimates of each Ψ_a, based on the n individuals sampled into phase-two, are

$$\hat{\Psi}_a = \frac{1}{N} \sum_{k=1}^{n} \frac{1}{\pi_n(V_k, Y_k)} \bar{Q}_n^{a*}(A = a, W_k, X_k). \tag{3.2}$$

Then, the IPCW-TMLE estimate of a contrast of marginalized risks, $h(\Psi_1, \Psi_0)$, is $\hat{\theta}_n = h(\hat{\Psi}_1, \hat{\Psi}_0)$. The influence-curve based variance of $\hat{\Psi}_a$ can be calculated as the sample variance of the n values $\pi_n(V_k, Y_k)^{-1} \bar{Q}_n^{a*}(A = a, W_k, X_k)$, and the delta method applied to calculate

the variance of $\hat{\theta}_n$. The fact that the IPCW-TMLE estimate of Ψ_a includes data from all n individuals from both exposure groups $a = 0, 1$, whereas the BH estimate only includes data from the subset with exposure level a, helps explain the potential for IPCW-TMLE to improve efficiency over BH.

3.3 Simulation Study Comparing Approaches

3.3.1 Design of Simulation Study

We conduct a simulation study with data sets generated to span a spectrum of design characteristics, including (a) a rare versus common endpoint (2.5% versus 25% failure rate), (b) few versus many endpoints (40 versus 400), and (c) dimensionality of the covariate data (5 versus 10 input variables). The performance of the IPCW-TMLE and BH methods is compared for each of the eight scenarios defined by the cross-classification of (a) $\times$ (b) $\times$ (c).

To simulate a data set, first an initial complete phase one data set comprised of enough participants N to obtain the specified number of cases is generated (e.g., for 400 cases and 25% failure rate N = 1,600, and for 400 cases and 2.5% failure rate N = 16,000). The complete data set includes a phase 1 covariate (V), a phase 1 binary outcome (Y), a phase 2 exposure (A), and phase 2 covariates ($W_1 - W_9$). In practice, the phase 2 covariates and binary exposure variable are only measured for the phase 2 sample; however, all variables are generated for all participants so that the two-phase methods can be compared to the full data version of the methods (i.e., TMLE and logistic regression) applied to the full data set. The covariate V used for phase 2 selection is generated with a Bernoulli distribution with success probability 0.5. The phase 2 covariates $W_1 - W_9$ are generated as independent continuous covariates with $W_l \sim Normal(0, 1)$, $l = 1, \ldots, 9$. To generate the binary exposure A of interest, we first generate a continuous exposure $A^c = -4 - 4W_4 + 4W_4^2$, and then set $A = 1$ if $A^c > 0$ and $A = 0$ if $A^c \leq 0$.

The outcome is generated as $Y \sim Bernoulli(p)$ where p is calculated as

$$p^{25\%} = expit[-5 - 2A^c - W_1^2 - sinW_1^2 + 3AsinW_1^2 +$$

$$15log(W_2^2) - W_3 - 2log(W_2^2)(1 - W_3)]$$

and

$$p^{2.5\%} = expit[-5 - 0.2A^c - 5W_1^2 - 5sinW_1^2 + 0.3AsinW_1^2 +$$

$$6log(W_2^2) - W_3 - 2log(W_2^2)(1 - W_3)]$$

for the simulations in which the outcome failure rate is 25% and 2.5%, respectively. This complex model for the outcome creates a scenario in which the truth cannot be correctly guessed and modeled with either the BH or IPCW-TMLE approaches, which puts the double robustness property of IPCW-TMLE to the test.

For each combination of factors explored, 200 full data sets are generated from which phase 2 samples are drawn via stratified Bernoulli sampling. All cases are selected for phase 2; controls are selected by Bernoulli sampling with selection probability $P(\Delta = 1|Y = 0, V) = 0.667$ for failure rate 25% and $P(\Delta = 1|Y = 0, V) = 0.051$ for failure rate 2.5%, such that we expect twice the number of controls sampled as cases for each of the two V strata.

Misspecified BH models are estimated as simple main-effects models and include more terms than are actually associated with the outcome: for the "5 covariate" misspecified model, the logistic regression includes A, V, and $W_1 - W_4$ as main effect terms; for the "10 covariate" misspecified model, the logistic regression includes A, V, and $W_1 - W_9$ as main effect terms. IPCW-TMLE is applied such that the SuperLearner library of estimators of each $\bar{Q}_a$ only considers untransformed V and $W_1 - W_4$ (or $W_1 - W_9$), such that the true outcome regression model is not included in the specified library of learners. For comparison, the BH and IPCW-TMLE methods are also implemented using the correctly specified models

as described above.

The simulation study uses the Superlearner R package for the step of estimating each $\bar{Q}_a$. We use a library of 9 estimators/algorithms to obtain the Superlearner initial estimate $\bar{Q}_n^a$, specifically SL.mean, SL.step, SL.bayesglm, SL.glm, SL.gam, SL.gam.3, SL.gam.4, SL.glm.interaction, and SL.glmnet. All are included in the SuperLearner package except SL.gam.3 and SL.gam.4 which are slight modifications of the SL.gam function. The SL.mean function fits a weighted sample mean, SL. step performs a stepwise selection logistic regression optimizing AIC, SL.bayesglm fits a bayesian logistic regression, SL.glm fits a standard logistic regression, SL.gam fits a generalized additive model with 2 degrees of smoothing while SL.gam.3 and SL.gam.4 use 3 and 4 degrees of smoothing. SL.glm.interaction fits a logistic regression with all 2-way interactions and SL.glmnet fits a LASSO model selecting λ via cross validation such that error is within one standard error of the minimum. Additionally, the non-negative least squares ("method.NNLS") loss function is used for all implementations of SuperLearner. The same Superlearner procedure used for estimation of each $\bar{Q}_a$ is used for obtaining the estimate $g_n(A = 1|W, X)$ of $g_0(A = 1|W, X)$. The two-phase sampling probabilities $\pi(V_k, Y_k)$ are estimated for each of the four categories $(v, y) \in \{0, 1\} \times \{0, 1\}$ by the four empirical frequencies with phase-two numerators and phase-one denominators.

BH and IPCW-TMLE estimates of each marginalized mean parameter of interest $\Psi_a = E[E(Y|A = a, W, X)]$, and variances of these estimates, are computed for each simulated data set. Performance of the estimators is evaluated by bias (Monte Carlo average of estimates $\hat{\Psi}_a$ minus true value Ψ_a), Monte Carlo standard error (MCSE) (square-root of sample variance of the $\hat{\Psi}_a$ estimates) as compared to the average estimated standard errors (SEs) of $\hat{\Psi}_a$, and root mean square error (RMSE). In addition, individual-level classification accuracy of the estimated models $\hat{E}[Y|A = a, W, X]$ fit as part of each method is evaluated by validated area under the ROC curve (AUC) estimated as the average across 200 additionally generated validation data sets (from the same true data generating distributions) of size $N = 10{,}000$.

We also compare BH and IPCW-TMLE estimation and inference about relative marginalized risk $\theta = \Psi_1/\Psi_0$. The delta method is applied to compute the variance of $log(\hat{\Psi}_1/\hat{\Psi}_0)$ and hence 95% confidence intervals about this parameter (and a Wald test of $log(\theta) = 0$); the confidence limits are then exponentially transformed to obtain 95% confidence intervals for θ.

3.4 Results

Table 3.1 displays estimated validation AUC for the IPCW-TMLE and BH methods fit to phase 2 data sets, as well as for TMLE and logistic regression (standard GLM) fit to the full (phase 1) data sets for comparison. AUC for IPCW-TMLE ranges from 0.93 to 0.98, while for BH predictive accuracy is much lower with AUC ranging from 0.59 to 0.74. AUC values for TMLE and for GLM are similar to the AUC values for IPCW-TMLE and BH, respectively. Therefore the two-phase sampling design appears to effective at maximizing prediction accuracy without needing to measure the phase-two data in a much larger set of control participants.

Table 3.2 reports percent bias and coverage probabilities of 95% confidence intervals (with Monte Carlo average confidence interval widths). All methods generally show small bias, albeit for scenarios with a small number of cases bias can be substantial. The BH and GLM methods exhibit greater bias than the IPCW-TMLE and TMLE methods, especially in estimation of Ψ_1 for the incidence rate of 25%. For both methods, most coverage probabilities of 95% confidence intervals for Ψ_1 and Ψ_0 fall above 90%. IPCW-TMLE sometimes has coverage probabilities greater than 95%, generally not showing undercoverage, whereas for some scenarios the BH and GLM confidence intervals undercover, especially for Ψ_1 under the common (25%) failure rate. IPCW-TMLE is often conservative, and this result is to be expected when the missingness probability is estimated.

Table 3.3 displays MCSE and average estimated standard errors (SE). For both IPCW-

TMLE and BH methods, the MCSE values are similar to or slightly lower than the average estimated standard errors, indicating that the standard error estimates appear to be reasonably accurate. The MCSE values for the IPCW-TMLE and BH approaches are also very similar to one another, indicating similar levels of efficiency for the two approaches. The MCSE estimates for the both the IPCW-TMLE and TMLE approaches are slightly larger for most scenarios than for the corresponding BH or GLM approaches. This may be explained by the fact that the BH and GLM approaches use the W and V information in their models as a form of covariate-adjustment.

Table 3.4 displays the relative RMSEs, calculated either as $RMSE^{BH}/RMSE^{GLM}$ or $RMSE^{IPCW-TMLE}/RMSE^{GLM}$, allowing assessment of the relative performance of each two-phase method compared to a standard logistic regression using the complete data set. The results for $RMSE^{BH}/RMSE^{GLM}$ range from 1.01 to 1.13, showing only slight efficiency loss by using a sub-sampling design. The RMSE ratios for the IPCW-TMLE approach are often slightly larger than those for the BH approach, which we think may stem from the results of Table 3.3 indicating that the BH Wald-based confidence intervals sometimes have too-small coverage probabilities, especially for the large sample size setting.

For IPCW-TMLE and BH estimation and inference on relative marginalized risks $\theta = \Psi_1/\Psi_0$, Table 3.5 displays percent bias, 95% confidence interval coverage probabilities, and 95% confidence interval widths for scenarios in which there were 400 cases. We see similar amounts of bias across methods except for the 25% failure rate scenario in which case the IPCW-TMLE methods have less bias than the BH methods, which is driven by the greater bias of $\hat{\Psi}_1$.

Table 3.1: Estimated validation data set AUC averaged over 200 validation data sets for the 8 failure rate, case count, and covariate number combinations. For phase 2 data analysis, the BH method with a misspecified logistic regression model and the IPCW-TMLE method with a misspecified Superlearner model are used to estimate the predicted probabilities of failure. For complete data analysis (before sampling phase 2 variables), the same misspecified logistic regression model (standard GLM) and TMLE method with the same misspecified Superlearner model are used to estimate the predicted values of failure.

	Validated AUC			
	Phase 2 Data		Complete Data	
%, Cases, Vars	IPCW-TMLE	BH	TMLE	GLM
25%, 40, 5	0.97	0.72	0.97	0.72
25%, 400, 5	0.98	0.74	0.99	0.74
25%, 40, 10	0.97	0.72	0.97	0.72
25%, 400, 10	0.98	0.74	0.98	0.74
2.5%, 40, 5	0.96	0.62	0.97	0.61
2.5%, 400, 5	0.97	0.63	0.97	0.63
2.5%, 40, 10	0.93	0.59	0.95	0.59
2.5%, 400, 10	0.97	0.63	0.97	0.63

Table 3.2: Percent biases and 95% confidence interval (CI) coverage probabilities for estimation and inference on Ψ_1 and Ψ_0. True values of parameters as estimated from a simulated data set of size $N = 10^9$ are $\Psi_0^{2.5\%} = 0.04$, $\Psi_0^{25\%} = 0.39$, $\Psi_1^{2.5\%} = 0.017$, and $\Psi_1^{25\%} = 0.06$.

	Phase 2 Data				Complete Data			
	IPCW-TMLE		BH		TMLE		GLM	
%, Cases, Var	% Bias	Coverage (95% CI width)	% Bias	Coverage (95% CI width)	% Bias	Coverage (95% CI width)	% Bias	Coverage (95% CI width)
$\Psi_1 = E[E(Y\|A=1,W,X)]$								
25%, 40, 5	65.5	1.00 (0.16)	63.9	0.92 (0.13)	65.5	1.00 (0.16)	62.6	0.94 (0.13)
25%, 400, 5	-0.6	0.96 (0.04)	13.8	0.84 (0.03)	-0.6	0.96 (0.04)	13.7	0.84 (0.03)
25%, 40, 10	13.2	0.99 (0.10)	24.0	0.92 (0.08)	13.3	0.99 (0.10)	23.2	0.92 (0.08)
25%, 400, 10	-0.6	0.96 (0.04)	13.7	0.84 (0.03)	-0.6	0.96 (0.04)	13.6	0.84 (0.03)
2.5%, 40, 5	1.8	0.96 (0.02)	1.0	0.94 (0.01)	1.2	0.94 (0.02)	1.0	0.92 (0.01)
2.5%, 400, 5	-0.2	0.93 (0.01)	-0.7	0.92 (0.01)	-0.6	0.92 (0.01)	0.1	0.94 (0.01)
2.5%, 40, 10	7.4	0.99 (0.03)	4.4	0.96 (0.02)	7.6	0.98 (0.02)	5.4	0.96 (0.02)
2.5%, 400 , 10	-1.6	0.92 (0.01)	-1.7	0.91 (0.01)	-2.0	0.91 (0.01)	-0.9	0.93 (0.01)
$\Psi_0 = E[E(Y\|A=0,W,X)]$								
25%, 40, 5	1.2	0.94 (0.19)	2.3	0.94 (0.18)	1.2	0.93 (0.18)	1.5	0.92 (0.17)
25%, 400, 5	0.3	0.96 (0.06)	-0.7	0.94 (0.06)	0.2	0.96 (0.06)	-0.7	0.94 (0.06)
25%, 40, 10	1.0	0.98 (0.14)	0.8	0.96 (0.12)	1.0	0.96 (0.13)	0.3	0.94 (0.12)
25%, 400, 10	0.3	0.96 (0.06)	-0.7	0.94 (0.06)	0.3	0.95 (0.06)	-0.7	0.94 (0.06)
2.5%, 40, 5	0.2	0.98 (0.02)	1.0	0.98 (0.02)	0.0	0.96 (0.02)	0.5	0.96 (0.02)
2.5%, 400, 5	-0.2	0.96 (0.01)	0.3	0.94 (0.01)	-0.3	0.90 (0.01)	0.1	0.90 (0.01)
2.5%, 40, 10	-2.3	0.98 (0.03)	-0.6	0.96 (0.03)	-2.3	0.92 (0.02)	-1.7	0.93 (0.02)
2.5%, 400, 10	0.0	0.97 (0.01)	0.4	0.95 (0.01)	-0.1	0.91 (0.01)	0.3	0.91 (0.01)

Table 3.3: MCSE and SE results for estimation and inference on Ψ_1 and Ψ_0 for the 8 failure rate, case count, and covariate number combinations. Estimates are presented as $\times 1000$. True values of parameters as estimated from a simulated data set of size $N = 10^9$ are $\Psi_0^{2.5\%} = 0.04$, $\Psi_0^{25\%} = 0.39$, $\Psi_1^{2.5\%} = 0.017$, and $\Psi_1^{25\%} = 0.06$.

	2 Phase				Complete Data			
	IPCW-TMLE		BH		TMLE		GLM	
%, Cases, Var	SE	MCSE	SE	MCSE	SE	MCSE	SE	MCSE
$\Psi_1 = E[E(Y\|A=1,W,X)]$								
25%, 40, 5	41.9	23.6	33.2	26.3	41.0	23.6	32.1	25.3
25%, 400, 5	10.5	10.6	8.7	9.8	10.8	10.7	8.6	9.6
25%, 40, 10	24.8	19.7	20.3	19.0	24.5	19.7	19.8	18.7
25%, 400, 10	10.5	10.6	8.7	9.8	10.4	10.6	8.6	9.6
2.5%, 40, 5	4.5	4.1	3.6	3.7	4.0	4.1	3.3	3.4
2.5%, 400, 5	2.0	1.8	1.5	1.7	1.8	1.8	1.5	1.5
2.5%, 40, 10	6.7	5.3	5.5	5.0	5.8	5.3	4.7	4.6
2.5%, 400, 10	1.9	1.8	1.5	1.7	1.8	1.8	1.5	1.5
$\Psi_0 = E[E(Y\|A=0,W,X)]$								
25%, 40, 5	49.3	50.0	45.2	48.4	47.2	50.0	44.0	49.3
25%, 400 , 5	15.6	14.9	14.5	14.4	15.0	15.7	14.1	14.3
25%, 40, 10	34.9	31.5	31.8	31.1	33.2	31.5	30.9	30.9
25%, 400, 10	15.6	14.9	14.4	14.4	14.8	14.9	14.0	14.3
2.5%, 40, 5	5.2	3.9	4.6	4.0	4.2	3.9	4.0	3.9
2.5%, 400, 5	2.3	2.0	2.0	2.0	1.9	2.0	1.8	2.0
2.5%, 40, 10	7.4	5.9	6.9	6.2	5.8	5.9	5.6	5.7
2.5%, 400, 10	2.3	2.0	2.0	2.0	1.9	2.0	1.8	2.0

Table 3.4: Estimated relative (IPCW-TMLE or BH/GLM) RMSE for estimation and inference on Ψ_1 and Ψ_0 for the 8 failure rate, case count, and covariate number combinations.

	Relative RMSE	
%, Cases, Var	IPCW-TMLE	BH
$\Psi_1 = E[E(Y\|A=1,W,X)]$		
25%, 40, 5	1.12	1.03
25%, 400, 5	0.98	1.02
25%, 40, 10	1.07	1.02
25%, 400, 10	0.98	1.02
2.5%, 40, 5	1.29	1.09
2.5%, 400, 5	1.26	1.07
2.5%, 40, 10	1.30	1.13
2.5%, 400, 10	1.25	1.08
$\Psi_0 = E[E(Y\|A=0,W,X)]$		
25%, 40, 5	1.06	1.01
25%, 400, 5	1.07	1.02
25%, 40, 10	1.08	1.02
25%, 400, 10	1.07	1.02
2.5%, 40, 5	1.17	1.09
2.5%, 400, 5	1.15	1.06
2.5%, 40, 10	1.18	1.16
2.5%, 400, 10	1.15	1.06

Table 3.5: Estimated relative marginalized risk (RR) percent bias and 95% confidence interval coverage (Monte Carlo averages) for scenarios with 400 cases. True RR, as estimated from a data set of size $n = 10^9$, is $RR^{2.5\%} = 0.42$ and $RR^{25\%} = 0.16$.

	IPCW-TMLE		BH	
		Coverage		Coverage
%, Var	%Bias	(95% CI width)	%Bias	(95% CI width)
25%, 5	-0.75	0.88 (0.08)	14.65	0.79 (0.09)
25%, 10	-0.76	0.88 (0.08)	14.53	0.79 (0.09)
2.5%, 5	0.20	0.92 (0.19)	-0.76	0.94 (0.18)
2.5%, 10	-1.41	0.89 (0.18)	-1.92	0.95 (0.18)

3.5 Application to Dengue and HIV-1 Vaccine Efficacy Trials

3.5.1 CYD14 & CYD15: Dengue Vaccine Efficacy Trials

Two randomized, double-blinded, placebo-controlled, multicenter, Phase 3 trials of the identical recombinant, live, attenuated, tetravalent dengue vaccine (CYD-TDV) versus placebo were conducted in Asia [9]) and Latin America [70]). These two trials–labeled CYD14 and CYD15–randomized 10,275 and 20,869 children, respectively, in 2:1 allocation to vaccine:placebo, with immunizations administered at months 0, 6, and 12. The primary analyses assessed vaccine efficacy (VE) against symptomatic, virologically confirmed dengue (VCD) occurring at least 28 days after the third immunization through to the Month 25 visit. Based on a proportional hazards model, estimated VE was 56.5% (95% CI 43.8–66.4) for CYD14 and 64.7% (95% CI 58.7–69.8) for CYD15. Sanofi Pasteur conducted the CYD14 and CYD15 studies and provided access to the study data (ClinicalTrials.gov identifiers NCT01373281 and NCT01374516, respectively).

The trials measured neutralizing antibody titers to each of the four dengue serotypes contained in the CYD-TDV vaccine at baseline and at Month 13. These titers were measured via case-cohort sampling performed as Bernoulli random samples of all randomized participants at enrollment and additionally from all participants who experienced the VCD endpoint after Month 13 and by Month 25 (cases). An individual's Month 13 $\log_{10}$-transformed geometric mean titer to the four vaccine-strain serotypes ("M13 average titer") has been studied as a biomarker of dengue risk and vaccine efficacy in secondary analyses [69, 40, 58], and we study this marker as our phase 2 variable of interest.

We apply the BH and IPCW-TMLE methodologies to assess, in the vaccine group, the association of M13 average titer with the subsequent risk of VCD through Month 25. We investigate how the risk of VCD compares between vaccine recipients with highest versus lowest tertile values of M13 average titer, coded $A = 1$ and $A = 0$ respectively, where

individuals with middle tertile values are excluded from the analysis. Our parameters of interest are the two marginalized risks $\Psi_1 = E[E(Y = 1|A = 1, X)]$ and $\Psi_0 = E[E(Y = 1|A = 0, X)]$, as well as the relative marginalized risk $\theta = \Psi_1/\Psi_0$, where X represents the phase 1 baseline covariates age, gender, and country of origin. BH and IPCW-TMLE estimates of Ψ_1, Ψ_0, and θ are obtained using the methods described in Sections 3.2.3 and 3.2.4, for the latter using the same Superlearner models for obtaining $\bar{Q}_n^1$ and $\bar{Q}_n^0$ and for obtaining $g_n(A = 1|X)$, and empirical frequencies are used for $\pi_n(y)$ for $y = 0, 1$. (There is no stratification variable V to condition on.) Variance estimates of $\hat{\Psi}_1$ and $\hat{\Psi}_0$ are computed as described in earlier sections, and used to calculate Wald 95% confidence intervals and Wald 2-sided p-values. Wald-based inference for $\theta = \Psi_1/\Psi_0$ is also done as described above.

Estimates of θ, the relative marginalized risk of VCD for top tertile M13 average titer participants to bottom tertile M13 average titer participants, are 0.062 and 0.076 for the BH and IPCW-TMLE approaches, respectively (Table 3.6). The 95% confidence intervals and Wald p-values indicate strong evidence for a lower risk of VCD among vaccine recipients with highest tertile M13 average titer responses compared to those with lowest tertile responses. These results support M13 average titer as a strong correlate of risk of dengue disease over the subsequent year of follow-up.

3.5.2 HVTN 505 HIV-1 Vaccine Efficacy Trial

To address the need for a safe and effective vaccine for the prevention of human immunodeficiency virus type 1 (HIV-1), in 2009-2013 a DNA prime and recombinant adenovirus type 5 boost (DNA/rAd5) vaccine regimen was administered to persons at increased risk for HIV-1 infection in the United States in a double-blind, placebo-controlled, HIV-1 vaccine efficacy trial [27]. At 21 study sites, 2504 men and transgender women who have sex with men were randomly assigned to receive the DNA/rAd5 vaccine (1253 participants) or placebo (1251 participants), in four injections at Month 0, 1, 3, 6. The primary analysis investigated the

Table 3.6: Estimation and inference on the marginalized risks of virologically confirmed dengue disease (VCD) for third tertile and first tertile M13 average titer[1] subgroups of vaccine recipients in the CYD14 & CYD15 studies combined

Model[2]	Relative Risk θ	95% CI	p-value
BH	0.062	(0.036, 0.11)	<0.001
IPCW-TMLE	0.076	(0.048, 0.12)	<0.001
	Risk Upper Tertile Ψ_1	95% CI	
BH	0.0023	(0.0013, 0.0034)	
IPCW-TMLE	0.0030	(0.0023, 0.0037)	
	Risk Lower Tertile Ψ_0	95% CI	
BH	0.037	(0.031, 0.044)	
IPCW-TMLE	0.039	(0.030, 0.048)	

[1] Log10 geometric mean of the four Month 13 neutralizing antibody titers to the four CYD-TDV dengue vaccine strains
[2] BH logistic regression model predicts VCD based on age, gender, country, and upper versus lower tertile indicator for M13 average titer. IPCW-TMLE Superlearner model considers the same input variables.

primary endpoint HIV-1 infection diagnosed from Month 7 to the final visit at Month 24. The study was completed early in April 2013 when a group sequential boundary for lack of vaccine efficacy was reached. Secondary objectives assessed several post-vaccination immune response biomarkers measured from Month 7 blood samples as correlates of risk of HIV-1 infection through Month 24; these biomarkers are ADCP, FcR2aConSgp140, IgA Env, IgG3 Env breadth, and CD8 Env PFS (polyfunctionality score), which are described in the clinical papers Janes et al. [32], Fong et al. [17], Neidich et al. [42], and Li et al. [36].

The secondary objective is assessed in vaccine recipients based on a two-phase sampling design, which measured the five Month 7 biomarkers from all cases (vaccine recipients who acquired HIV-1 infection between Month 7 and 24) and a Bernoulli stratified random sample of 125 controls (vaccine recipients who reached the Month 24 visit without diagnosis of HIV-1 infection). The variable V for stratified sampling was the cross-classification of BMI category (18.4–25.0; 25.0–29.8; 29.8–40.0) and race/ethnicity (white, black, Hispanic). The BH and IPCW-TMLE methods are applied as described above, with phase 1 (baseline) covariates

including age, race/ethnicity, BMI, and a baseline behavioral risk variable used in previous analyses (the papers cited above). The Month 7 immune response biomarkers are modeled as binary indicators for above ($A = 1$) versus at or below ($A = 0$) the median marker value. The methods were applied separately to each of the five immune response biomarkers. In addition, we studied the IgA Env and ADCP biomarkers together, by considering the four subgroups defined by above versus at or below the medians of these biomarkers, and assessing relative risks of three of these subgroup versus the (Low, Low) reference subgroup.

Table 3.7 displays results on estimation and inference on the relative marginalized risks of HIV-1 infection. The biomarkers ADCP, CD8 Env PFS, and IgG3 Env breadth are significantly associated with HIV-1 risk, with above-median response associated with about 4-fold lower risk than below-median response. There is not evidence that FcR2aConSgp140 or IgA Env associate with risk of HIV-1 infection (all p-values > 0.20). There is some evidence that vaccine recipients with Low IgA Env and High ADCP have lower HIV-1 risk compared to vaccine recipients with Low IgA Env and Low ADCP, with stronger relative risk than vaccine recipients with High IgA Env and High ADCP. When comparing the estimates obtained by the two methods (BH and IPCW-TMLE), the estimates for relative risk for all immune response markers are similar with substantially overlapping confidence intervals. The results suggest similar or slightly greater precision (narrower confidence intervals) for estimation using IPCW-TMLE compared to BH.

Table 3.7: Estimation and inference on relative marginalized risks of HIV-1 infection for Month 7 immune response biomarkers in vaccine recipients of the HVTN 505 study. Each immune response biomarker is studied separately (together with the phase 1 variables).

Immune Response Biomarker[1]	Model[2]	Relative Risk[3]	95% CI	p-value
ADCP	BH	0.28	(0.09, 0.88)	0.029
	IPCW-TMLE	0.34	(0.13, 0.90)	0.030
CD8 Env PFS	BH	0.20	(0.05, 0.74)	0.016
	IPCW-TMLE	0.22	(0.08, 0.59)	0.003
IgG3 Env breadth	BH	0.19	(0.04, 0.87)	0.032
	IPCW-TMLE	0.23	(0.07, 0.69)	0.009
FcR2aConSgp140	BH	0.80	(0.35, 1.83)	0.593
	IPCW-TMLE	0.72	(0.30, 1.70)	0.452
IgA Env	BH	1.15	(0.55, 2.40)	0.714
	IPCW-TMLE	1.20	(0.51, 2.83)	0.681
IgA Env high; ADCP high[4]	BH	0.43	(0.14, 1.33)	0.142
	IPCW-TMLE	0.44	(0.17, 1.17)	0.099
IgA Env high; ADCP low[4]	BH	1.10	(0.40, 3.07)	0.851
	IPCW-TMLE	1.30	(0.49, 3.40)	0.598
IgA Env low; ADCP high[4]	BH	0.11	(0.01, 1.04)	0.054
	IPCW-TMLE	0.15	(0.02, 1.08)	0.060

[1] Binary indicator for above-median immune response value.

[2] BH logistic regression model predicts HIV-1 based on age, race/ethnicity, BMI, a behavioral risk variable (Hammer et al. (2013)), and the specified immune response biomarker of interest. IPCW-TMLE Superlearner model considers the same input variables.

[3] Relative maginalized risk of HIV-1 infection for participants with Month 7 immune response biomarkers above the median compared to those below.

[4] "High" indicates above median value and "low" indicates below median value for the respective immune response biomarker. Relative risk is comparison of the specified immune response biomarker combination indicator relative to risk for low IgA and low ADCP group.

3.6 Discussion

In the simulation study, both the well-established Breslow-Holubkov logistic regression based method and the more recently developed IPCW-TMLE method for data analysis of two-phase sampling studies performed similarly in some scenarios, however there were differences in classification accuracy, bias, confidence interval coverage, and consistency between the methods in other scenarios. The Superlearner model used as an intermediate step of IPCW-TMLE had greater classification accuracy than the logistic regression model used as part of the BH approach. Additionally, in some scenarios the BH approach exhibited a greater magnitude of bias for estimation of Ψ_1, and had lower confidence interval coverage (under-coverage) for estimation of the relative marginalized risk, implying that in some circumstances the IPCW-TMLE approach out-performs the BH approach. In application to two dengue and one HIV-1 vaccine efficacy clinical trial, the BH and IPCW-TMLE approaches yielded similar estimates of exposure-specific infection or disease risks and of marginalized relative risks, with the IPCW-TMLE confidence intervals being wider for the HIV-1 study. Further research evaluating performance of IPCW-TMLE and BH for phase 2 stratification variables (V) with more than two levels and for other types of data generating distributions could expand the applicability of these types of simulated comparisons. Overall, the IPCW-TMLE approach shows promise in achieving results with desirable classification and inferential accuracy, thus providing a desirable option for analysis in addition to the well-established method of Breslow and Holubkov. Example code demonstrating the methods in this manuscript is available as an R package at https://github.com/brendalewisprice/IPCWTMLE2Phase.

3.7 Calculation of Standard Errors

3.7.1 Derivation of Standard Errors for Predicted Values from a Logistic Regression Using iid Observations

We start by defining $X = n$ x p as a matrix of n observations with p covariates, and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ as the vector of coefficients in a logistic regression that seeks to model those p variables to predict a binary outcome Y. The first step in estimating the predicted values and their standard errors is to estimate the distribution of $\hat{\beta}$: $\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, \Sigma)$. For a model fit with `glm` in R, the estimates for β and the covariance matrix for β, $\Sigma = \Sigma_{glm}$, are standard output from the `glm` function. For the BH approach, the needed estimates ($\hat{\beta}$ and $\hat{\Sigma}_{BH}$) are also output with the model fit using the `osDesign` package [28]). As application of the delta method for both approaches (glm or BH) is the same, going forward this illustration will only display the estimates derived from an application of the BH approach. To obtain standard errors for the glm approach, one would only have to substitute Σ_{glm} for Σ_{BH} and use the corresponding glm model estimate of β.

Using these estimates and the delta method, one can easily estimate the distribution of $\hat{X\beta} = logit(\hat{p})$, the logit of the predicted values for a set of n observations, using the relationship

$$\sqrt{n}(\hat{X\beta} - X\beta) \rightarrow N(0, X^T \Sigma_{BH} X)$$

where $X^T \Sigma_{BH} X$ is the n x n covariance matrix of the individual observations, and $\hat{\Sigma}_{BH}$ would be used to estimate the covariance matrix.

For application of the delta method in estimating standard errors for the predicted values $\hat{p}$ instead, one would need to apply the expit function ($\hat{p} = g(\beta; x) = expit(x\hat{\beta}) = e^{x\hat{\beta}}/(1 + e^{x\hat{\beta}})$). There are readily available packages in R that can estimate these values, depending on the model used, including a "deltamethod" function in the `msm` package [31] that correctly

calculates the standard errors for $X\beta$ and for the $expit(X\beta)$ transformation. Below we illustrate in more detail the steps needed for this application of the delta method that results in estimates of individual level p and functions of p.

To obtain estimates for a single predicted $logit(\hat{p})$ value we use the distribution

$$\sqrt{n}(x\hat{\beta} - x\beta) \quad \rightarrow N(0, x\Sigma_{BH}x^T).$$

Therefore, for a single $\hat{p} = g(\hat{\beta}; x) = expit(x\hat{\beta}) = \frac{e^{x\hat{\beta}}}{1+e^{x\hat{\beta}}}$ we apply the transformation $g(\beta; x)$ and use the following estimates and gradient in application of the delta method.

$$\begin{aligned} g(\beta; x) &= expit(x\beta) = \frac{e^{x\beta}}{1 + e^{x\beta}} \\ \nabla g(\beta; x) &= (\frac{\partial g}{\partial \beta_1} \dots \frac{\partial g}{\partial \beta_p})^T \\ \frac{\partial g}{\partial \beta_k} &= \frac{x_k e^{x\beta}}{(1 + e^{x\beta})^2} \end{aligned}$$

where $k = 1 \dots p$. Our distribution for a single predicted value can be written as

$$\sqrt{n}(expit(x\hat{\beta}) - expit(x\beta)) \quad \rightarrow N(0, \nabla g(\beta; x)^T \Sigma_{BH} \nabla g(\beta; x))$$

and therefore for the full data set as

$$\sqrt{n}(expit(X\hat{\beta}) - expit(X\beta)) \quad \rightarrow N(0, \nabla g(\beta; X)^T \Sigma_{BH} \nabla g(\beta; X))$$

where

$$\begin{aligned}\nabla g(\beta; X) &= \begin{bmatrix} \frac{\partial g^{x1}}{\partial \beta_1} \cdots \frac{\partial g^{x1}}{\partial \beta_p} \\ \vdots \ddots \vdots \\ \frac{\partial g^{xn}}{\partial \beta_1} \cdots \frac{\partial g^{xn}}{\partial \beta_p} \end{bmatrix} \\ \frac{\partial g^{xi}}{\partial \beta_k} &= \frac{x_k(e^{xi\beta}(1+e^{xi\beta})-e^{2xi\beta})}{(1+e^{xi\beta})^2}\end{aligned}$$

For estimation of the standard errors, we use the estimated $\hat{\beta}$s for the βs in the gradients, and the estimated $\hat{\Sigma}_{BH}$ for the covariance matrix.

The variance for a single iid predicted value is simple:

$$var_{g(x_i,\beta)} := \nabla g(\beta; x_i)^T \Sigma_{BH} \nabla g(\beta; x_i).$$

For other functions of the data, $h(x, y)$, an additional application of the delta method would be needed.

Treatment or Exposure Group Mean

In the previous section we described how to obtain the estimated distribution for predicted values from a BH logistic regression:

$$\sqrt{n}(expit(X\hat{\beta}) - expit(X\beta)) \quad \rightarrow N(0, \nabla g(\beta; X)^T \Sigma_{BH} \nabla g(\beta; X))$$

An additional application of the delta method can be used to derive the asymptotic distribution of functions ($h(x, y)$) of these individual predicted values. For a single treatment group weighted mean (with weights $w_i \ldots w_{n_1}$),

$$\widehat{E}[E(Y|A = 1, X)] = \frac{1}{\sum_{j=1}^{n_1} \hat{w}_j} \sum_{A_i=1} \widehat{w}_i E(Y|A = 1, X_i)$$

we could use the delta method with the following gradient:

$$\frac{\partial h(g(\beta;X)}{\partial g(\beta;X)} = (\frac{\hat{w}_1}{\sum_{j=1}^{n_1}\hat{w}_j}, \ldots, \frac{\hat{w}_{n_1}}{\sum_{j=1}^{n_1}\hat{w}_j}, 0_1, \ldots, 0_{n_0})^T$$

So, for the weighted variance estimates, based on the phase 2 or observation weights ($w_i = \frac{\Delta_i}{\pi_n(V,Y)}$), the weight is incorporated into the gradient.

From there the variance is estimated as

$$Var_n(\widehat{E}[E(Y|A=1,X)]) = \hat{\sigma}^2_{EY_1} = \frac{\partial h(g(\beta;X))}{\partial g(\beta;X)}^T \nabla g(\beta;X)^T \Sigma_{BH} \nabla g(\beta;X) \frac{\partial h(g(\beta;X))}{\partial g(\beta;X)}$$

Calculation of the standard error follows directly from there, calculated as

$$SE_n(\widehat{E}[E(Y|A=1,X)]) = \sqrt{\frac{\hat{\sigma}^2_{EY_1}}{\sum_{j=1}^{n_1}\hat{w}_j}}$$

For another example, say one wants to calculate the estimated weighted average exposure effect (AEE):

$$AEE \;=\; \frac{1}{\sum_{j=1}^{n_1}\hat{w}_j}\sum_{A_i=1}^{n_1}\hat{w}_i\widehat{E}(Y|A=1,X_i) - \frac{1}{\sum_{j=1}^{n_0}\hat{w}_j}\sum_{A_i=0}^{n_0}\hat{w}_i\widehat{E}(Y|A=0,X_i).$$

All that is needed is to derive and apply the appropriate gradient, which would be

$$\frac{\partial h(g(\beta;X))}{\partial g(\beta;X)} = \left(\frac{w_{1,1}}{\sum_{j=1}^{n_1}w_j}, \ldots, \frac{w_{1,n_1}}{\sum_{j=1}^{n_1}w_j}, -\frac{w_{0,1}}{\sum_{j=1}^{n_0}w_j}, \ldots, -\frac{w_{0,n_0}}{\sum_{j=1}^{n_0}w_j}\right)^T$$

.

3.7.2 Calculation of Standard Errors for IPCW-TMLE

For a single unweighted treatment group mean,

$$\widehat{E}[E(Y|A=1,X)] \;=\; \frac{1}{n}\sum_{A_i=1}\widehat{E}(Y|A=1,X_i)$$

the influence curve can be derived as

$$IC_1 = \frac{I(A=1)}{g_0(1|X)}(Y - E_0(Y|X,A=1) + \bar{Q}_{1,0}(X) - \psi_0(1)$$

or, more generally

$$IC_a = \frac{I(A=a)}{g_0(a|X)}(Y - E_0(Y|X,A=a) + \bar{Q}_{a,0}(X) - \psi_0(a).$$

The weighted treatment group mean for treatment $A=1$, where the individual weight $w_i = \frac{\Delta_i}{\pi_n(V,Y)}$, can be computed as

$$\widehat{E}[E(Y|A=1,X)] \;=\; \frac{1}{\sum_{j=1}^{n}\hat{w}_j}\sum_{i=1}^{n}\hat{w}_i\hat{E}(Y|A=1,X_i)],$$

and the influence curve for the weighted treatment group mean is expressed as

$$IC_1 = \frac{\Delta}{\pi_n(V,Y)}\frac{I(A=1)}{g_0(1|X)}(Y - E_0(Y|X,A=1)) + \frac{\Delta}{\pi_n(V,Y)}(\bar{Q}_{1,0}(X) - \psi_0(1)),$$

or, more generally

$$IC_a = \frac{\Delta}{\pi_n(V,Y)}\frac{I(A=a)}{g_0(a|X)}(Y - E_0(Y|X,A=a)) + \frac{\Delta}{\pi_n(V,Y)}(\bar{Q}_{a,0}(X) - \psi_0(a)).$$

Calculation of the standard error for the weighted treatment mean follows easily from

these influence curves [59]. Under TMLE the mean influence curve equals 0, $I\bar{C}_a = 0$, so the variance of $\widehat{E}[E(Y|A,X)]$ is calculated as

$$Var(\widehat{E}[E(Y|A=a,X)]) = \frac{1}{\sum_{j=1}^{n}\hat{w}_j} var(IC_a) = \frac{1}{\sum_{j=1}^{n}\hat{w}_j}(IC_a - I\bar{C}_a)^2$$

And the standard error follows naturally as

$$SE_a = \sqrt{\frac{IC_{n,a}^2}{\left(\sum_{j=1}^{n}\hat{w}_j\right)^2}},$$

with estimated values of IC_a and w_j used in the calculation for estimated standard error.

Calculation of Standard Errors for Relative Risk using IPCW-TMLE

Estimation of 95% confidence intervals is best done on the log relative risk (RR) scale and then back-transformed to the RR scale for interpretation. The influence curve for the weighted log relative risk can be written as

$$\begin{aligned} IC_{logRR} \quad &= \tfrac{1}{\mu_1}\left(\tfrac{\Delta}{\pi_n(V,Y)}\tfrac{I(A=1)}{g_0(1|X)}(Y - E_0(Y|X,A=1) + \tfrac{\Delta}{\pi_n(V,Y)}(\bar{Q}_{1,0}(X))\right) \\ &-\tfrac{1}{\mu_0}\left(\tfrac{\Delta}{\pi_n(V,Y)}\tfrac{I(A=0)}{g_0(0|X)}(Y - E_0(Y|X,A=0) + \tfrac{\Delta}{\pi_n(V,Y)}(\bar{Q}_{0,0}(X))\right) \end{aligned}$$

where μ_1 and μ_0 are the weighted treatment means and can be estimated as $\hat{\mu}_1 = \widehat{E}[E(Y|A=1,X)]$ and $\hat{\mu}_0 = \widehat{E}[E(Y|A=0,X)]$.

The standard error for the log relative risk follows naturally as

$$SE_{logRR} = \sqrt{\frac{IC_{n,logRR}^2}{\left(\sum_{j=1}^{n}\hat{w}_j\right)^2}}$$

from which confidence intervals on the log scale can be constructed and then back-transformed.

Chapter 4

IPCW-TMLE METHODOLOGY FOR TWO-PHASE STUDIES WITH OUTCOMES SUBJECT TO RIGHT-CENSORING

4.1 Introduction

When Rose and van der Laan presented IPCW in two-phase case-control studies they briefly commented on how it could be applied to right-censored data (scenarios in which the failure time is subject to right-censoring)[55]. In this chapter we elaborate on their suggested approach, presenting directly how it would be executed. TMLE has been developed and applied to two-phase studies and case-control studies [55, 65] and has also been developed for some time-to-event scenarios [65, 66]. Application of IPCW-TMLE to scenarios with two-phase data in which the failure time is subject to right-censoring is an important approach for real data applications, with preventive vaccine efficacy trials being one example. This chapter extends the work of Rose and van der Laan by applying the IPCW-TMLE approach to two-phase case-control studies with right censoring.

In this scenario, the full-data structure is a right-censored data structure from which we conduct a two-phase study. Right-censoring refers to the feature of the data that some subjects will be lost to follow-up before the event occurs, or, at the end of the study, a subject has not failed yet and is subject to administrative censoring. In particular, we are considering scenarios in which participants are followed over time for the occurrence of the event $Y = 1$ during a fixed follow up period running through to a fixed time τ, where there are K discrete visits $t_1, \ldots, t_K$ for some $k \in \{1, \cdots, K\}$. Therefore, the indicator $Y = I(T \leq t)$ represents the occurrence of the outcome of interest, where $t = t_k$ refers to a single timepoint of interest. We consider the right-censored data structure for a two-

phase sampling design in which the first stage is done by taking a sample from the target population and vector of information V is gathered on each individual. In this first phase, the researcher measures $\tilde{T}$, where $\tilde{T} = min(T, C)$, the minimum of the time the event of interest T and the censoring variable C. The variable C is recorded as the time when a subject is no longer enrolled in the study, potentially due to drop out or administrative censoring, with $\Delta = I(\tilde{T} \leq C)$ representing the failure indicator. We assume conditionally independent censoring ($C \perp\!\!\!\perp T|(X_1, A)$), indicating a nonzero probability of not being censored through the end of the observation period $P(C \geq \tau|X_1) > 0$. We define $A \in \{0, 1\}$ as the treatment assignment at baseline or other binary exposure of interest measured at baseline, X_1 denotes the baseline vector of covariates, and W denotes the vector of covariates measured later on during phase 2 (discussed below). If A is a marker such as an immune response, this set-up applies with baseline defined as the time of the marker measurement.

The second phase then involves taking a subsample from this original sample and collecting additional information on the selected subsample. Our observed outcome, $Y^* = (\tilde{T} \leq t, \Delta = 1)$ is the indicator of having observed failure by the fixed time t. We include an indicator $R = 1$ for denoting membership in the phase 2 sample, therefore our observed data structure is given by $O = (R, RW, RA, X_1, V, \tilde{T}, \Delta, Y)$, where W represents additional covariates collected on the phase 2 sample, and X can be used to represent the full data structure $X = (X_1, V, W, A, \tilde{T}, \Delta, Y)$. Therefore, the observed data includes n i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P_0$. The true probability distribution of X can be denoted by $P_{X,0}$, with $\mathcal{M}^F$ representing a statistical model for $P_{X,0}$. The target parameter of the full data distribution $\Psi^F : \mathcal{M}^F \to \mathbb{R}^d$ gives us $\psi_0^F = \Psi^F(P_{X,0})$ as our parameter of the true probability distribution of X. Similar to before, the efficient influence curve of Ψ^F for the full-data distribution of P_X will be written as $D^F(P_X)$. The two-phase sampling weights are derived as described in Section 2.1.1 and the true probability of sampling to phase two is denoted in similar fashion as $\Pi_{R,0} \equiv P_{X,0}(R = 1|X)$, with $\Pi_{R,n}$ representing an estimate of the probability

of sampling. In this section, as noted above, since Δ is used for our right-censoring indicator, R now represents the indicator for sampling, such that the two-phase sampling weights will now be written as $R/\Pi_{R,0}$, which differs slightly from the notation used in Chapter 3.

It is important to note that the definition of a case in a two-phase case-cohort study with this type of right-censoring is slightly different than in a two-phase case-cohort study without right-censoring. In a typical cumulative case-control study within a prospective cohort study, when there is no right-censoring we define cases as those for which an outcome of interest has been observed and controls as participants who completed follow-up without the outcome observed. In the setting of two-phase case control data with right censoring, we have two more aspects of the data to take into account: (1) missing data on Y due to both loss to follow up prior to observation of T or due to administrative censoring, and (2) missing data on Y due to the way in which controls are defined as a subsample of the "eligible control" population. The variable defining a case, $Y^* = 1$ could be defined, say, as observed death by month t, $Y^* = I(\tilde{T} \leq t, \Delta = 1)$; the "eligible controls," could then be defined as those subjects for which $\{\tilde{T} = \tau, \Delta = 0\}$ (observed to be administratively censored at end of follow-up τ), and controls for analysis would then be taken as a sample from all eligible controls. The sample would typically be either Bernoulli, stratified Bernoulli, sampling without replacement, or a stratified without replacement sample. Those samples that are unstratified would be case-control samples, and those that are stratified samples would be two-phase samples. In this chapter we restrict to Bernoulli random sampling that may be stratified. The Bernoulli proportion for control assignment would be set to $\Pi_{R,n}(\tilde{T} = \tau, \Delta = 0)$, the proportion selected for the second stage sampling, and $\Pi_{R,n}(Y^* = 1)$ would be equal to 1. However Y^* is not our outcome of interest. Table 4.1 illustrates the difference in what is observed and the question of interest that might occur in these types of scenarios. We see that for cases, $Y^* = 1$ implies $Y = 1$ but not vice versa, and, for controls, $(\tilde{T} = \tau, \Delta = 0)$ implies $Y = 0$ but not vice versa.

For both cases and controls, the sampled cohorts are subsets of the cohorts that are representative of the populations of interest. The actual outcome by t for those that are censored prior to observing their event and prior to t is unknown or missing. This missing data problem is often not directly addressed, but dealt with by omitting those censored observations (potentially introducing bias), or by including right-censored individuals as eligible controls. Previous work has explored TMLE approaches that take into account right censoring [41, 60, 65] and missing data structures [49, 52]. Additionally, Rose and van der Laan [56] touched upon the combination of both but did not elaborate. The approach presented here elaborates on the application to both conditions concurrently as part of the missing data problem in which estimation of nuisance parameters is used to adjust for the missingness, and thus, under appropriate assumptions, can result in consistent and asymptotically normal estimates of the quantities of interest.

Table 4.1: Motivation: Populations of Interest vs Sampled

	Of Interest	$\neq$	Sampled
Cases	$Y = I(T \leq t) = 1$		$\tilde{T} \leq t, \Delta = 1$
Controls	$Y = I(T \leq t) = 0$		$\tilde{T} = \tau, \Delta = 0$

4.2 Parameters of Interest

The parameters of interest discussed in this chapter are similar to those discussed in Chapter 3, Section 3.2.2. As before, we can define two statistical parameters of interest from which a variety of contrasts can be constructed to answer questions of interest. For clarity, let T_1 represent a subject's time to event if the subject had been assigned treatment $A = 1$ and T_0 represent the subject's time to event had they been assigned treatment $A = 0$. T_1 and T_0 are the counterfactual time to event values for each subject, regardless of which counterfactual

was actually observed in the study. For these data, the conditional hazard is given by $\lambda(\cdot|A,W)$ with the corresponding conditional survival function represented as $S(\cdot|A,W)$. Observation is continued through a fixed follow up time τ over K discrete visits $t_1,\ldots,t_K$ where t_k represents a particular time point of follow-up of interest, that is, one of the K time points $t_1,\cdots,t_K$. In this setting with time-to-event data we define the parameters of interest as

$$\psi_1(t_k) = P(T_1 > t_k) = E_0[S_0(t_k|A=1,W,X)] \tag{4.1}$$

for treatment group $A=1$ and

$$\psi_0(t_k) = P(T_0 > t_k) = E_0[S_0(t_k|A=0,W,X)] \tag{4.2}$$

for treatment group $A=0$, where t is the timepoint of interest. We refer to ψ_a as the marginalized exposure-specific mean for each $a=\{0,1\}$. Then, for any continuous and differentiable contrast function $h(x,y)$ satisfying $h(x,y)=0$ if and only if $x=y$ and $h(x,y)<0$ for $x<y$, one can estimate a contrast parameter $\theta = h(\psi_1,\psi_0)$ by $\hat{\theta} = h(\hat{\psi}_1,\hat{\psi}_0)$, where $\hat{\psi}_a$ is an estimate of ψ_a. In this chapter we will focus on estimating ψ_a, however, as presented before, three common quantities of interest using this approach include $h(x,y)=x-y$, the marginalized risk difference (RD); $h(x,y)=(1-x)/(1-y)$, the relative marginalized risk (RR); and $h(x,y)=[(1-x)/x]\,/\,[(1-y)/y]$, the marginalized odds ratio (OR).

$$\theta_{RD}(p_0)(t_k) = P(T_1 > t_k) - P(T_0 > t_k) \tag{4.3}$$

$$\theta_{RR}(p_0)(t_k) = \frac{1-P(T_1 > t_k)}{1-P(T_0 > t_k)} \tag{4.4}$$

$$\theta_{OR}(p_0)(t_k) = \frac{(1 - P(T_1 > t_k))/P(T_1 > t_k)}{(1 - P(T_0 > t_k))/P(T_0 > t_k)} \tag{4.5}$$

Furthermore, given estimates of $var(\psi_1)$ and $var(\psi_0)$, application of the delta method yields the variance of $\hat{\theta}$ for any contrast function $h(x, y)$.

4.3 Starting with the Full Data TMLE

To apply IPCW-TMLE to a two-phase study of time-to-event data with right censoring, we will begin with the full data approach, to which we can then apply the methods for two-phase studies. For this approach, the full data is that data set of time-to-event data, which includes some right censoring, but has not been subsampled for phase two. Additionally, the time to event data is such that a subject is observed over K fixed timepoints defined as $t_1, \cdots, t_K$ with $t = t_k$ the fixed time-point of interest. We present the targeted maximum likelihood estimation of a marginal treatment specific survival for a fixed end point, as was presented by Moore and van der Laan [41].

Begin with an initial fit $\hat{p}^0$ of the density on the observed data O. This is a fit determined by an initial hazard fit $\hat{\lambda}(t|A, W)$, the distribution of A (identified by $\hat{g}(A = 1|W)$ and $\hat{g}(A = 0|W) = 1 - \hat{g}(A = 1|W)$ for a binary A), the estimated censoring mechanism denoted by $\bar{G}(t|A, W)$, and the marginal distribution of W estimated by the empirical probability distribution of $W_1, \ldots, W_n$. In a randomized clinical trial (RCT), treatment is randomized such that $\hat{g}(A = 1|W) = \frac{1}{n}\sum_{i=1}^{n} A_i$, however, $\hat{g}(A = 1|W)$ can also be estimated using SuperLearner, which has been shown to improve efficiency [64].

For this approach we assume the survival time is discrete and we estimate the initial fit of the conditional hazard $\hat{\lambda}(t|A, W)$ using a logistic regression model of the form

$$logit(\lambda(t|A, W)) = \alpha(t) + m(A, W|\beta), \tag{4.6}$$

where α represent an intercept and $m(\cdot)$ is a parametric form of β. This is the first step in

the TMLE algorithm. From there, a selected working submodel is iteratively fit to update the estimate of $\lambda(t|A,W)$. In this case, the parametric working submodel would be

$$logit(\lambda(t|A,W)) = \alpha(t) + m(A,W|\beta) + \epsilon h(t,A,W) \tag{4.7}$$

in which the term $h(t,A,W)$ is added to the equation (4.6) by the TMLE approach for the estimation procedure. It becomes our parametric working model for fluctuating the conditional probability distribution of $\lambda(t|A,W)$ based on a log-likelihood loss function. The function $h(t,A,W)$ is specific to the parameter of interest and is selected such that the score for this hazard model in equation (4.7) at $\epsilon = 0$ equals or spans the full-data efficient influence curve under the assumption of coarsening at random (CAR). For the parameters of interest ψ_a, the marginalized exposure-specific means, these "clever covariates" $h_1(t,A,W)$ and $h_0(t,A,W)$ functions can be written as

$$h_a(t,A,W) = -\frac{I(A=a)}{g(a|W)\bar{G}(t|A,W)}\frac{S(t_k|A,W)}{S(t|A,W)}I(t \leq t_k) \text{ for } a \in \{1,0\} \tag{4.8}$$

for each $t = t_1,\ldots,t_k$, for $a \in \{1,0\}$, and then iteratively fit and updated until $\epsilon \approx 0$ in equation (4.7).

In practice, one finds an initial estimate of ϵ by fitting a logistic regression using the covariates of $m(A,W|\beta)$ and $h(t,A,W)$ with an initial fit on the hazard. The coefficient for $m(A,W|\beta)$ is fixed at one and the intercept α is set to zero such that only ϵ is estimated for the initial iteration of the targeted maximum likelihood algorithm. For further iterations, the updated estimate $\hat{\lambda}^1(t|A,W)$ is used in place of the initial fit and the covariate $h(t,A,W)$ is re-evaluated using updated $\hat{S}^1(t|A,W)$ based on the new estimate $\hat{\lambda}^1(t|A,W)$. From that fit is obtained $\hat{\lambda}^2(t|A,W)$, which is used for the third iteration in like manner as $\hat{\lambda}^1(t|A,W)$ was used for the second. Iteration continues until the estimate $\hat{\epsilon}$ is essentially zero, and the resulting targeted maximum likelihood estimate of λ from the final update, denoted by

$\hat{\lambda}^*(t|A,W)$ results in the corresponding survival TMLE from the final update, $\hat{S}^*(t|A,W)$. This process can be done separately for each ψ_a, but can also be done simultaneously for both by incorporating respective ϵ and $h(\cdot)$ terms for each parameter of interest in the working submodel:

$$logit(\lambda(t|A,W)) = \alpha(t) + m(A,W|\beta) + \epsilon_1 h_1(t,A,W) + \epsilon_0 h_0(t,A,W). \tag{4.9}$$

Iteration continues until both $\hat{\epsilon}_1$ and $\hat{\epsilon}_0$ are essentially zero. Once these TMLE estimates $\hat{S}^*(t|A,W)$ are obtained, estimates of the parameters of interest can be computed:

$$\hat{\psi}_a^*(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{S}^*(t_k|a,W_i) \text{ for } a \in \{1,0\}. \tag{4.10}$$

We assume that phase 2 variables are missing at random (MAR) (see [57, 55] for further discussion), which can be expressed by $P_{X,0}(R=1|X) = g_{R,0}(1|X) = g_{R,0}(1|V)$. That is, the probability of selection to phase two is assumed to be independent of any phase two covariates. Additionally, we assume conditionally independent censoring ($C \perp\!\!\!\perp T|(W,A)$), a nonzero probability of not being censored through the end of the observation period $P(C \geq \tau|A=a,W=w) > 0$ for each possible value of (w,a), ignorability stated as A is randomized conditional on X_1 and V, and positivity of the form $P(A=1|W=w) > 0$ and $P(A=0|W=w) > 0$ for each possible value of w.

In this full data approach for right censored data, we see that only the initial hazard estimate is updated at each iteration. While assuming MAR, the density of the observed data factorizes into the marginal distribution of the covariates W (p_W), the treatment mechanism $g_{A,0}(A|W)$, the conditional probability of censoring $\bar{G}(t|A,W)$, and the product over time of the conditional hazard $\lambda(t|A,W)$. This implies that the components of the efficient influence curve corresponding to $g_{A,0}(A|W)$ and $\bar{G}(t|A,W)$ are zero, leaving only the components p_w and $\lambda(t|A,W)$. By using the empirical distribution of W to estimate p_w we are using the

nonparametric maximum likelihood estimate for p_w and therefore do not need to update it at each iteration. This leaves only the hazard estimate, $\hat{\lambda}(t|A,W)$, needing updating at each iteration of the algorithm.

The efficient influence curves for the respective parameters of interest, ψ_1 and ψ_0, can be written as

$$D_a(p_0) = \sum_{t \leq t_k} h_a(g, \bar{G}, S)(t, A, W) \left[I(\tilde{T} = t, \Delta = 1) - I(\tilde{T} \geq t)\lambda(t|A = a, W) \right] + S(t_k|A = a, W) - \psi_a(p_0)(t_k) \text{ for } a \in \{1, 0\} \tag{4.11}$$

where the first term is the sum over t up to t_k of the "clever covariate" $h_a(\cdot)$ multiplied by the residual. One of the advantageous properties of this approach is that if either the survival function $S(\cdot|A,W)$ (and thus $\lambda(\cdot|A,W)$) is consistently estimated, or if both the treatment and censoring mechanisms $g_A(A|W)$ and $\bar{G}(\cdot|A,W)$ are consistently estimated, then the TMLE estimates will be consistent. However, if the survival function is consistently estimated, but neither the treatment nor censoring mechanisms are, the estimate could not be considered consistently estimated. For example, assume that the treatment mechanism is correctly specified, as can be guaranteed by randomization (discussed further below). Then, double robustness leaves us needing only the survival or censoring mechanisms consistently estimated. Or, for another example, assume there is either no censoring or the censoring mechanism is consistently estimated. Then we only need either the survival or treatment assignment consistently estimated. Furthermore, if all conditions are met, then $\hat{\psi}_1^*(t)$ is also efficient. In randomized clinical trials (RCTs), the treatment is often assigned, and thus the treatment mechanism is known such that $g_A(A|W)$ can certainly be estimated consistently, which then gives us one of the conditions "for free," thus focusing concern on consistent estimation to $\bar{G}(\cdot|A,W)$ or $S(\cdot|A,W)$ for RCTs. If one observes no censoring, or if one can assume censoring is missing at random and $\bar{G}(\cdot|A,W)$ is correctly modeled, then the TMLE

will be consistent, and, if $S(\cdot|A,W)$ is consistently estimated, also efficient.

One can utilize the estimate of the efficient influence curve (equation (4.11)) to construct Wald-type confidence intervals for a parameter of interest. The asymptotic variance for $\sqrt{n}(\hat{\psi}_1^*(t) - \psi_1(t))$ can be obtained with the efficient influence curve

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} D_1^2(\hat{p}^*)(O_i). \tag{4.12}$$

Subsequent hypothesis tests for those estimated parameters of interest can be executed using this estimate of σ^2. For example, to test the null hypothesis $H_0 : \Psi_1(p_0)(t_k) \leq M$ for some constant M reflecting a survival probability of interest, one can use the test statistic

$$T_n = \frac{\hat{\psi}_1^*(t) - M}{\hat{\sigma}/\sqrt{N}}. \tag{4.13}$$

Additionally, to test the null hypothesis $H_0 : \Psi_1(p_0)(t) - \Psi_0(p_0)(t) = 0$ one can use the test statistic

$$T_n = \frac{\hat{\psi}_1^*(t) - \hat{\psi}_0^*(t)}{\hat{\sigma}/\sqrt{N}} \tag{4.14}$$

where the estimate of $\hat{\sigma}$ is obtained using the delta method with the respective treatment group estimates of $\hat{\sigma}$.

Application of the delta method to the estimates of σ for the respective parameters of interest can be used to estimate standard errors for additional contrasts of interest.

4.4 Applying IPCW-TMLE to the Full Data TMLE

Now that we have established the TMLE approach for right-censored data, we can apply the IPCW-TMLE approach to that full data structure. Remember, our observed data structure consists of n i.i.d. copies of $O_1, \ldots, O_n$, where $O = (V, R, RX)$. In the previous section we

detailed the TMLE approach for the full data structure that does not include sub-sampling based on on R. Assuming missing at random (MAR), we have $g_{R,0}(r|X) = g_{R,0}(r|V)$, and, once again, for notational convenience going forward we will let $\Pi_0(V) \equiv g_{R,0}(r = 1|V)$. To apply the IPCW-TMLE approach to the full data structure, we will include weights of $R_i/\Pi_n(V_i)$ for all observations $i = 1, \ldots, n$, where, as specified above, we use $\Pi_n(V)$ to refer to our estimator of $g_{R,0}(r = 1|V)$. So, applying this approach to the full data structure, we begin by applying the observation level weights to our initial estimator of $\lambda(t|A, W)$. In this manner we are approaching the initial fit as an IPCW-loss based fit. Referring back to our description of the full data approach, this entails fitting the initial hazard in like manner as done in equation (4.7), however, now using a weighted logistic regression model of the form of equation (4.7) with observation level weights applied. Thus, the initial estimate is adjusted with the IPCW-TMLE weights to account for the missingness structure imposed by the two-phase sampling. We begin our iterative process by fitting a weighted logistic regression like was done in equation (4.7):

$$logit(\lambda(t|A, W)) = \alpha(t) + m(A, W|\beta) + \epsilon h(t, A, W) \tag{4.15}$$

where now h_a incorporates the two-phase weights for the regression:

$$h_a(t, A, W) = -\frac{R}{\Pi_n(V)} \frac{I(A = a)}{g(a|W)\bar{G}(t|A, W)} \frac{S(t_k|A, W)}{S(t|A, W)} I(t \leq t_k). \tag{4.16}$$

For iterations $m = 2, \ldots, M$, we can set the terms in equation (4.15) to 1 for $\alpha(t)$ and $\hat{\lambda}^{m-1}(t|A, W)$ for $m(A, W|\beta)$ for each m^{th} iteration such that the model solves for ϵ on each subsequent iteration. Iteration continues until ϵ_a is nearly zero (or in the case of equation (4.9), both ϵ_0 and ϵ_1 are nearly zero).

Once these IPCW-TMLE estimates $\hat{S}^*(t|A, W)$ are obtained, estimates of the parameters

of interest can be computed using a weighted mean:

$$\hat{\psi}_a^*(t) = \frac{1}{\sum_{i=1}^N \frac{R_i}{\Pi_n(V_i)}} \sum_{i=1}^N \frac{R_i}{\Pi_n(V_i)} \hat{S}^*(t|a, W_i) \tag{4.17}$$

The efficient influence curve can now be written as

$$\begin{aligned} D_a(p_0) = &\sum_{t \leq t_k} h_a(g_0, \bar{G}_0, S_0, R)(t, A, W)[I(\tilde{T} = t, \Delta = 1, R = 1) \\ &- I(\tilde{T} \geq t, R = 1)\lambda_0(t|A = a, W)] \\ &+ \frac{R}{\Pi_0(V)} \left[S_0(t_k|A = a, W) - \psi_a(p_0)(t_k)\right]. \end{aligned} \tag{4.18}$$

And an estimate of the variance of $\hat{\psi}_a^*(p^*)(t)$, for use in confidence intervals and hypothesis tests, can be obtained using the efficient influence curve:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N D_1^2(\hat{p}^*)(O_i). \tag{4.19}$$

Though the above description and example is for a targeted maximum likelihood estimator, as also discussed in Rose and van der Laan [55], this presentation of IPCW-TMLE is not only a targeted maximum likelihood estimator, but also a targeted minimum loss based estimator with a loss function defined as an IPCW full-data loss function and using a parametric submodel $P_X(\epsilon)$ with a weighted score function $R/\Pi_0(V)D^F(P_X(\epsilon))$ at $\epsilon = 0$. Additionally, double robust properties of the full-data efficient influence curve are continued in the IPCW-TMLE. This process can be described more generally using the notation of $L^F(P_X)(X)$ for the full data loss function. In this IPCW-TMLE application, the initial estimator is instead fit with $\frac{R}{\Pi_n(V)} L^F(P_X)(X)$. And then for the iterative process of $1, \ldots, K$ iterations, for the full data structure we would have $\epsilon_n^k = argmin_\epsilon \frac{1}{n} \sum_{i=1}^n L^F(P_{X,n}^{k-1}(\epsilon))(X_i)$ so for the IPCW-TMLE application we now have $\epsilon_n^k = argmin_\epsilon \frac{1}{n} \sum_{i=1}^n \frac{R_i}{\Pi_n(V_i)} L^F(P_{X,n}^{k-1}(\epsilon))(X_i)$.

The identifiability of the parameters of interest as causal parameters, where $\psi_a = E[Y_a]$, is additionally dependent on assumptions that were also mentioned in Chapter 3, Section 3.2.2. We make the common causal assumptions of (i) SUTVA, (ii) Consistency, (iii) Ignorability (i.e., A is randomized conditional on W), and (iv) Positivity (i.e. $P(A = 1|W = w) > 0$ a.e. and $P(A = 0|W = w) > 0$ a.e.), as well as $\bar{G}(t|A, W) \equiv \prod_{t=t_1}^{t_k}(1 - \bar{g}(t|A, W)) > 0$ a.e., which was addressed in previous work by Robins and Rotnitzy [54]).

4.4.1 Statistical Properties

Section 4.3 discussed the double robust advantages of the full data (right censored) structure. We observed that if either the survival function $S(\cdot|A, W)$ (and thus $\lambda(\cdot|A, W)$) is consistently estimated for the full data, or if both the treatment and censoring mechanisms $g(A|W)$ and $\bar{G}(\cdot|A, W)$ are consistently estimated, the TMLE estimates on the full data will be consistent. Now with the two-phase sampling of the data, we also consider how estimation of $R/\Pi(V)$ plays a roll in consistency and efficiency of the TMLE estimates. The conditions now become that if either the survival function $S(\cdot|A, W)$ (and thus $\lambda_0(\cdot|A, W)$) is consistently estimated for the full data, or if the treatment, censoring, and two-phase sampling mechanisms $g(A|W)$, $\Pi(V)$, and $\bar{G}(\cdot|A, W)$ are consistently estimated, the TMLE estimates on the full data will be consistent. Given that the two-phase sampling mechanism is often chosen by the researcher, estimation of $\Pi(V)$ can often be considered consistent. So, in cases of RCT in which we can assume both $g(A|W)$ and $\Pi(V)$ are consistently estimated, if the estimation of $S(\cdot|A, W)$ or $\bar{G}(\cdot|A, W)$ is consistently estimated, then the IPCW-TMLE estimates will be asymptotically consistent, and if both $S(\cdot|A, W)$ and $\bar{G}(\cdot|A, W)$ are consistently estimated, the IPCW-TMLE estimates will be efficient.

4.5 SuperLearner for Predicted Values

SuperLearner is a flexible and nonparametric way in which to estimate the initial predicted values input into the TMLE algorithm. Below we discuss previous work on using SuperLearner for estimation of the initial predicted values for right-censored data and for IPCW-TMLE. Following we discuss how both can be easily executed together.

4.5.1 SuperLearner for Right-Censored Data

The first step of IPCW-TMLE to right censored two phase data requires estimating the initial hazard in the example outlined above, or predicted values for the parameter of interest. Application of the SuperLearner to right censored data is discussed by Polley and van der Laan in Chapter 16 of van der Laan and Rose's *Targeted Learning* [65]. They consider several possible parameters of interest for survival outcomes including a user-supplied function of T given the baseline covariates. In particular, they consider $\psi_0(W) = E_0(m(T)|W)$ where $m(T)$ could be $m(T) = T$, $m(T) = logT$, and $m(T) = I(T > t)$. The function $m(T) = I(T > t)$ is of particular interest for several reasons, including that it can yield the conditional survival function which is often a parameter of interest. The approach by Polley and van der Laan [65] for estimating $m(T) = I(T > t)$ and the conditional survival function considers an IPCW squared error loss function approach in which the weight is actually a function of the censoring variable Δ and a nuisance parameter of the conditional survival function of censoring time C given W, $\bar{G}_0(T|A,W) = P_0(C > t|A,W)$. In particular, they proposed using the squared error loss function

$$L\left(O,\psi_a\right) = \frac{\Delta}{\bar{G}_0(T|A,W)} \cdot \left[m(T_a) - \psi_a\right]^2. \tag{4.20}$$

If the censoring is known to be independent then the marginal survivor function $\bar{G}_0(t) =$

$P(C > t)$ can be estimated with the Kaplan-Meier estimator:

$$\bar{G}_{KM,n,a}(t) = \prod_{t \leq t_k} \left(1 - \frac{\sum_{i=1}^{n} I\left(\tilde{T}_i = t, \Delta_i = 0, A = a\right)}{\sum_{i=1}^{n} \left(\tilde{T}_i \geq t, A = a\right)} \right).$$

Although Polly didn't touch upon it, SuperLearner itself can be used to estimate the censoring probability $P(C > t|A, W)$ without making the assumption that censoring is independent of baseline characteristics or treatment; in fact, baseline covariates can be used to model the probability of censoring. SuperLearner, using baseline or phase one covariates available for all observations (W), can be used to estimate the probability of censoring conditional on those given baseline characteristics $P(C > t|A, W)$ using a library of potential learners and a pre-specified loss function without having to restrict to a particular estimator, like the Kaplan-Meier estimator.

Polley goes on to discuss two ways to weight the potential algorithms in the cross-validation selection of the optimal combination of algorithms such that the loss function is bounded and the oracle inequality results apply. For the IPCW loss function constraining the sum of weights to 0 with non-negative weights is one option, and potentially the most common option used:

$$\left\{ \sum_{k} \alpha_k \psi_k : \sum_{k} \alpha_k = 1, \alpha_k \geq 0 \right\}. \tag{4.21}$$

For the log-likelihood loss function, a similar approach is used however optimization is on the logit-scale transformed predicted values instead of the actual predicted values. Sometimes a symmetric truncated logit link is used to deal with conditional hazard values approaching 0 or 1 ([65], pp. 254-255). Since this chapter is interested in the combination of two-phase sampling and right censored data, we will focus on the IPCW loss function used for censoring.

4.5.2 SuperLearner with IPCW Addressing Two-phase Sampling

When applying SuperLearner to two-phase data, a similar approach that incorporates a type of weighting is applied. As touched upon in Chapter 3, Section 3.2.4, the SuperLearner uses a loss function that reflects the sub-sampling step used to select the phase-two data. For the IPCW squared error loss function that accounts for the phase-two sampling weights as defined in this chapter, $R/\Pi(V)$, we have

$$L\left(O,\psi\right) = \frac{R}{\Pi(V)} \cdot \left[m(T_a) - \psi_a\right]^2 \tag{4.22}$$

which is very similar in form to equation (4.20). For many loss functions, application of the individual observation level weights reflecting probability of sampling will be important. Weighting of potential algorithms in the cross-validation selection of the optimal combination of algorithms in order to ensure the oracle inequality applies is similar to the discussion in Section 4.5.1 and equation (4.21).

4.5.3 Combining Approaches for SuperLearner

For both equation (4.20) and equation (4.23)the loss function is weighted by a quantity that reflects a form of censoring: in equation (4.20) it reflects the right-censoring nature of the data; in equation (4.23), the missing data structure of a phase two sample. In both cases there is a nuisance parameter to be estimated in order to apply the loss function. In equation (4.20) it was the marginal survivor function $\bar{G}_0(t|A,W)$; in equation (4.23) it is the probability of sampling to the phase 2 sample. Both of these quantities can be estimated using SuperLearner, which tends to improve performance by accounting for finite-sample random imbalances in confounders [64]. The combined IPCW squared error loss function for

that setting can be written as

$$L\left(O,\psi\right)=\underbrace{\frac{R}{\Pi_0(V)}}_{\text{Sampling}}\cdot\underbrace{\frac{\Delta}{\bar{G}_0(T|A,W)}}_{\text{Censoring}}\cdot\underbrace{\left[m(T_a)-\psi_a\right]^2}_{\text{MSE Loss}}. \tag{4.23}$$

4.6 Discussion

This chapter detailed the approach for applying IPCW-TMLE to right-censored data structures in a way to target parameters of interest. This approach allows estimation of marginalized mean parameters like $E[E[I(T \leq t)|A = a, W]]$, as well as their contrasts, accounting for the fact that the outcome of interest $Y = I(T \leq t)$ is subject to right censoring. Applying IPCW-TMLE to a right-censored data approach allows one to target such quantities and obtain estimators that retain desirable properties such as asymptotic consistency and efficiency. Valid estimates of the standard error of IPCW-TMLE estimates of the parameters of interest can be obtained, thus allowing researchers to create asymptotically valid confidence intervals and hypothesis tests. Continuing research in this topic includes simulation of data similar to right-censored vaccine efficacy trials to determine characteristics of bias and efficiency in the approach, when compared to a full-data approach, and when compared to other methods commonly used.

Chapter 5
DISCUSSION

5.1 Summary

In this dissertation we explored the application of TMLE in estimation of an optimal surrogate and the implementation of inverse probability of censoring weighted targeted minimum loss-based estimation (IPCW-TMLE) to uncensored and right-censored two-phase data structures. In Chapter 1 we developed methodology for the estimation of optimal surrogates in randomized trials using targeted minimum loss-based estimation (TMLE), first in the setting of complete data, and then in Chapter 2, extended to the setting of two-phase data. This approach meets the proposed minimal requirement for an intermediate endpoint to be a useful surrogate endpoint in that it avoids the surrogate paradox. The newly proposed approach starts at this minimal requirement by defining the optimal surrogate in a way guaranteed to satisfy the Prentice definition of a valid surrogate and hence avoid the paradox. Furthermore, this approach to estimating an optimal surrogate uses unbiased supervised statistical learning and targeted minimum loss based estimation to estimate the optimal surrogate. This is advantageous in common applications where many baseline covariates and intermediate response endpoints are measured, but some uncertainty exists as to how best to predict the study outcome from these collected data. Moreover, while the focus was on randomized studies, we discussed how this framework will also apply for generating promising candidate surrogates based on observational studies when making a few additional assumptions. We considered both an ideal setting with no missing data and where the clinical outcome is never observed before the intermediate response endpoints are measured, and then a more common setting in which analysis is done on a two-phase sample of the same

data. Chapter 3 presented a comparison of IPCW-TMLE estimation to a commonly used method proposed by Breslow and Holubkov (BH) for parameter estimation in two-phase studies. The simulation study assessed the comparative differences in classification accuracy, bias, and efficiency of estimates obtained using both methods. The SuperLearner model used as an intermediate step of IPCW-TMLE had greater classification accuracy than the logistic regression model used as part of the BH approach. Additionally, in some scenarios the BH approach exhibited a greater magnitude of bias for estimation and had lower confidence interval coverage (under-coverage) for estimation of the relative marginalized risk, implying that in some circumstances the IPCW-TMLE approach out-performs the BH approach. In application to two dengue and one HIV-1 vaccine efficacy clinical trials, the BH and IPCW-TMLE approaches yielded similar estimates of exposure-specific infection or disease risks and of marginalized relative risks, with the IPCW-TMLE confidence intervals being wider for the HIV-1 study. Overall, the IPCW-TMLE approach showed promise in achieving results with desirable classification and inferential accuracy, thus providing a desirable option for analysis in addition to the well-established method of Breslow and Holubkov.

Chapter 4 elaborated on suggestions by Rose and van der Laan ([55]) that their two-phase IPCW-TMLE methodology for estimation could be applied for estimation of causal parameters of interest in right-censored two-phase studies. We outlined the full-data structure of right-censored time-to-event data upon which this would be based and then detailed the application of IPCW-TMLE to that full data structure in a way that the two-phase selection is thought of as a missing data problem that requires estimation of nuisance parameters. Assumptions needed for consistency and the asymptotically normal properties of the estimates for this setting was considered and described.

5.2 Limitations and Future Research

The methods developed in this dissertation have broad application to randomized clinical trials with two-phase designs for measuring biomarkers, however, further work is needed to explore potential real-world challenges that may occur in the application of these approaches. In Chapter 1, Theorems 2 and 3 describe conditions for using the optimal surrogate built from an efficacy trial to confer correct estimation of the clinical treatment effect in a new setting based on the estimated surrogate endpoint without measuring the clinical endpoint, in particular when the new setting entails the same treatment or a new treatment, respectively. Our applications to simulated data sets and to the actually clinical trial data indicated that additional research is needed to determine the effects of departures from the assumptions of randomization, equal conditional means, and contained support. Additionally, in order to provide confidence intervals for the clinical treatment effect in a new setting while also accounting for the error in estimating the optimal surrogate further research is needed, especially if the conditional mean is modeled nonparametrically.

In response to Chapter 3, further research evaluating the performance of IPCW-TMLE and BH for stratification variables (V) used for phase two selection when there are more than two levels for V and for other types of data generating distributions could expand the applicability of these types of simulated comparisons. Future work will also include expansion upon the work presented in Chapter 4. Additional simulation studies investigating the application of IPCW-TMLE to right-censored clinical trial data are planned in order to explore relative efficiency and bias in comparison to other methods.

BIBLIOGRAPHY

[1] A Alonso, G Molenberghs, H Geys, M Buyse, and T Vangeneugden. A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine*, 25:205–221, 2006.

[2] E Bareinboim and J Pearl. Transportability of causal effects: Completeness results. *Proceedings of the Twenty-Sixth National Conference on Artificial Intelligence, Menlo Park, CA*, pages 698–704, 2012.

[3] N Breslow, JH Lubin, and P Marek. Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78:1–12, 1983.

[4] NE Breslow and KC Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988.

[5] NE Breslow and R Holubkov. Maximum likelihood estimation of logistic regression parameters for two-phase outcome-dependent sampling. *J. Roy. Statist. Soc*, 59:447–461, 1997.

[6] NE Breslow, T Lumley, CM Ballantyne, LE Chambless, and M Kulich. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, 169:1398–1405, 2009.

[7] M Buyse and G Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014–1029, 1998.

[8] M Buyse, G Molenberghs, T Burzykowski, D Renard, and H Geys. The validation on surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1:49–67, 2000.

[9] MR Capeding, NH Tran, SRS Hadinegoro, Hussain Imam HJ Muhammad Ismail, Tawee Chotpitayasunondh, Mary Noreen Chua, Chan Quang Luong, Kusnandi Rusmil, Dewa Nyoman Wirawan, Revathy Nallusamy, et al. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *The Lancet*, 384(9951):1358–1365, 2014.

[10] ISF Chan, L Shu, H Matthews, C Chan, R Vessey, J Sadoff, and J Heyse. Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine*, 21:3411–3430, 2002.

[11] N Chatterjee, Y-H Chen, and NE Breslow. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98:158–167, 2003.

[12] H Chen, Z Geng, and J Jia. Criteria for surrogate end points. *Journal of the Royal Statistical Society, Series B*, 69:919–932, 2007.

[13] MJ Daniels and MD Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16:1965–1982, 1997.

[14] VL Ernster. Nested case-control studies. *Preventative Medicine*, 23(5):587–590, 1994.

[15] W Dana Flanders and Sander Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747, 1991.

[16] Y Fong and P B Gilbert. Calibration weighted estimation of semiparametric transformation models for two-phase sampling. *Statistics in Medicine*, 34(10):1695–707, 2015.

[17] Y Fong, X Shen, VC Ashley, A Deal, KE Seaton, C Yu, et al. 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. *Journal of Infectious Diseases*, 217(8):1280–1288, 2018.

[18] CE Frangakis and DB Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–29, 2002.

[19] LS Freedman, BI Graubard, and A Schatzkin. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11:167–178, 1992.

[20] E Gabriel and PB Gilbert. Evaluating principle surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics*, 15:251–265, 2014.

[21] MH Gail, R Pfeiffer, HC Van Houwelingen, and RJ Carroll. On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1:231–246, 2000.

[22] PB Gilbert, E Gabriel, Y Huang, and ISF Chan. Surrogate endpoint evaluation: Principal stratification criteria and the prentice definition. *Journal of Causal Inference*, 3(2):157–175, 2015.

[23] PB Gilbert and MG Hudgens. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64:1146–1154, 2008.

[24] PB Gilbert, MG Hudgens, and J Wolfson. Commentary on "Principal stratification– a goal or a tool?" by Judea Pearl. *The International Journal of Biostatistics*, 7:Article 1, 2011.

[25] PB Gilbert, L Qin, and SG Self. Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*, 27:4758–4778, 2008. PMCID: PMC2646675.

[26] E Hak, F Wei, DE Grobbee, and KL Nichol. A nested case-control study of influenza vaccination was a cost-effective alternative to a full cohort analysis. *Journal of Clinical Epidemiology*, 57(9):875–880, 2004.

[27] SM Hammer, ME Sobieszczyk, H Janes, ST Karuna, MJ Mulligan, et al. Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *New England Journal of Medicine*, 369:283–292, 2013. PMCID: NIHMS142560.

[28] S Haneuse, T Saegusa, and T Lumley. osdesign: an R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of Statistical Software*, 43, 2011.

[29] TJ Hastie and RJ Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, 1990.

[30] Y Huang, PB Gilbert, and J Wolfson. Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics*, 69:301–309, 2013.

[31] CH Jackson. Multi-state models for panel data: The msm for r. *Journal of Statistical Software*, 38(8):1–29.

[32] HE Janes, KW Cohen, N Frahm, SC De Rosa, et al. 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(9):1376–1385, 05 2017.

[33] MM Joffe. Surrogate measures and consistent surrogates discussion, 2013.

[34] MM Joffe and T Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65:530–538, 2009.

[35] F Kobayashi and M Kuroki. A new proportion measure of the treatment effect captured by candidate surrogate endpoints. *Statistics in Medicine*, 33:3338–3353, 2014.

[36] SS Li, PB Gilbert, LN Carpp, C Pyo, H Janes, Y Fong, et al. Fc gamma receptor polymorphisms modulated the vaccine effect on HIV-1 risk in the HVTN 505 HIV vaccine trial. *Journal of Virology*, pages JVI–02041, 2019.

[37] Y Li, JMG Taylor, and MR Elliott. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*, 66:523–531, 2010.

[38] DY Lin, TR Fleming, and V De Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, 16:1515–1527, 1997.

[39] N Mantel. Synthetic retrospective studies and related topics. *Biometrics*, 29(3):479–486, 1973.

[40] Z Moodie, M Juraska, Y Huang, Y Zhuang, Y Fong, LN Carpp, SG Self, L Chambonneau, R Small, N Jackson, F Noriega, and PB Gilbert. Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *Journal of Infectious Diseases*, 217(5):742–753, 2018.

[41] KL Moore and MJ van der Laan. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints.* Chapman and Hall/CRC Biostatistics Series, 2009.

[42] SD Neidich, Y Fong, S Li, D Geraghty, BD Williamson, WC Young, D Goodman, KE Seaton, X Shen, S Sawant, L Zhang, A deCamp, BS Blette, M Shao, NL Yates, F Feely, C-W Pyo, G Ferrari, I Frank, ST Karuna, E Swann, JM Mascola, BS Graham, SM Hammer, ME Sobieszczyk, L Corey, HE Janes, MJ McElrath, R Gottardo, PB Gilbert, and GD Tomaras. Antibody Fc effector functions and IgG3 associate with decreased HIV-1 risk. *Journal of Clinical Investigation*, in press, 2019.

[43] J Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33:101–116, 1938.

[44] J Pearl. *Direct and Indirect Effects.* Morgan Kaufmann, San Francisco, 2001.

[45] J Pearl. Discussion on ìsurrogate measures and consistent surrogatesî. *Biometrics*, 69(3):573–577, 2013.

[46] J Pearl and E Bareinboim. Transportability of causal and statistical relations: A formal approach. *Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence, Menlo Park, CA*, pages 247–254, 2011.

[47] E Polley, E LeDell, and MJ van der Laan. Superlearner: Super learner prediction. r package version 2.0-19. *https://CRAN.R-project.org/package=SuperLearner*, 2016.

[48] EC Polley and MJ van der Laan. Superlearner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, page Working Paper 226, 2010.

[49] KE Porter, S Gruber, MJ van der Laan, and JS Sekhon. The relative performance of targeted maximum likelihood estimators. *Int J Biostat*, 7(1), 2011.

[50] RL Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8:431–440, 1989.

[51] BL. Price, PB. Gilbert, and MJ van der Laan. Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, 74(4):1271–1281, 2018.

[52] JM Robins. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association: Section on Bayesian Statistical Science*, page 6–10.

[53] JM Robins and S Greenland. Identifiability and exchangeability of direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

[54] JM Robins, A Rotnitzky, and LP Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.

[55] S Rose and MJ van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*, 7(1):1–21, 2011. PMCID: PMC3083136.

[56] PR Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74:13–26, 1987.

[57] DB Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[58] S Sridhar, A Luedtke, E Langevin, M Zhu, M Bonaparte, T Machabert, S Savarino, B Zambrano, A Moureau, A Khromava, Z Moodie, T Westling, C Mascareñas, C Frago, M Cortés, D Chansinghakul, F Noriega, A Bouckenooghe, J Chen, S Ng, PB Gilbert, S Gurunathan, and CA DiazGranados. Effect of dengue serostatus on dengue vaccine safety and efficacy. *New England Journal of Medicine*, 4:327–340, 2018.

[59] LA Stefanski and DD Boos. The Calculus of M-Estimation The Calculus of M-Estimation. *The American Statistican*, 1(56):29–38, 2002.

[60] OM Stitelman and MJ van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6, Issue 1:Article 21, 2010.

[61] JMG Taylor, Y Wang, and R Thibaut. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61:1102–1111, 2005.

[62] MJ van der Laan, AE Hubbard, and SK Pajouh. Statistical inference for data adaptive target parameters. *U.C. Berkeley Division of Biostatistics Working Paper Series*, page Paper 314, 2013.

[63] MJ van der Laan, EC Polley, and AE Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[64] MJ Van der Laan and JM Robins. *Unified Methods for Censored Longitudinal Data and Causality.* Springer, 2003.

[65] MJ van der Laan and S Rose. *Targeted learning: Causal Inference for Observational and Experimental Data.* Springer, 2011.

[66] MJ van der Laan and S Rose. *Targeted learning in Data Science.* Springer, 2018.

[67] TJ VanderWeele. Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, 78:2957–2962, 2008.

[68] TJ VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69:561–568, 2013.

[69] C Vigne, M Dupuy, A Richetin, B Guy, N Jackson, M Bonaparte, B Hu, M Saville, D Chansinghakul, F Noriega, and E Plennevaux. Integrated immunogenicity analysis of a tetravalent dengue vaccine up to 4 y after vaccination. *Human vaccines & immunotherapeutics*, 13(9):2004–2016, 2017.

[70] L Villar, GH Dayan, JL Arredondo-García, DM Rivera, R Cunha, C Deseda, H Reynales, MS Costa, JO Morales-Ramírez, G Carrasquilla, et al. Efficacy of a tetravalent dengue vaccine in children in latin america. *New England Journal of Medicine*, 372:113–123, 2015.

[71] E Vittinghoff and DC Bauer. Case-only analysis of treatment–covariate interactions in clinical trials. *Biometrics*, 62(3):769–776, 2006.

[72] W Wang, D Scharfstein, Z Tan, and EJ MacKenzie. Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):947–969, 2009.

[73] Y Wang and JMG Taylor. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 58:803–812, 2002.

[74] CJ Weir and RJ Walley. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine*, 25:183–203, 2006.

[75] JE White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–28, 1982.

[76] G Yang, Y Sun, L Qi, and PB Gilbert. Estimation of stratified mark-specific proportional hazards models under two-phase sampling with application to HIV vaccine efficacy trials. *Statistics in Biosciences*, 9(1):259–283, 2017.