Published in final edited form as:
Biometrics. 2018 December ; 74(4): 1271–1281. doi:10.1111/biom.12879.

Estimation of the Optimal Surrogate Based on a Randomized Trial

Brenda L. Price[1], **Peter B. Gilbert**[1,2], and **Mark J. van der Laan**[3]

[1]Department of Biostatistics University of Washington, Seattle, Washington, 98109, U.S.A.

[2]Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, U.S.A.

[3]Division of Biostatistics, University of California, Berkeley, California, 94720, U.S.A.

Summary:

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 the setting that we observe n independent and identically distributed observations of a random variable consisting of baseline covariates, a treatment, a vector of candidate surrogate outcomes at an intermediate time point, and the final outcome of interest at a final time point. We assume the treatment is randomized, conditional on the baseline covariates. The goal is to use these data to learn a most-promising surrogate for use in future trials for inference about a mean contrast treatment effect on the final outcome. We define an optimal surrogate for the current study as the function of the data generating distribution collected by the intermediate time point that satisfies the Prentice definition of a valid surrogate endpoint and that optimally predicts the final outcome: this optimal surrogate is an unknown parameter. We show that this optimal surrogate is a conditional mean and present super-learner and targeted super-learner based estimators, whose predicted outcomes are used as the surrogate in applications. We demonstrate a number of desirable properties of this optimal surrogate and its estimators, and study the methodology in simulations and an application to dengue vaccine efficacy trials.

Keywords

Asymptotic linearity; Cross-validation; Efficient influence curve; Prentice definition of a valid surrogate; Semiparametric model; Super-learner; Targeted maximum likelihood; Targeted minimum loss based estimation

[8] Supplementary Materials
Web Appendices and Figures referenced in Sections 1 and 3–6 are available with this paper at the Biometrics website on Wiley Online Library: (A) 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; (B) A review of how the optimal surrogate framework compares to other surrogate evaluation frameworks; (C) A proof of Theorem 3; (D) the details of estimating the optimal surrogate via super-learner; (E)–(F) Expanded details of the Example (including on assumption diagnostics); (G) Two simulation studies of the proposed methodology; and (H) Dummy example data sets and R code producing all of the results for the dengue application (for the dummy data sets).

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 an 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 study as the function of the true data generating distribution collected by the intermediate time point that satisfies the Prentice definition of a valid surrogate endpoint and that optimally predicts the final outcome: this optimal surrogate is an unknown parameter. In Section 2 we show the highly desirable property that the optimal predictor automatically satisfies the Prentice definition, with one appealing consequence that this optimal surrogate guarantees avoidance of the disastrous ‘surrogate paradox’ [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] (VanderWeele, 2013) cannot occur. 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 3 we give conditions under which 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. This describes “transportability assumptions” under which the average treatment effect on the optimal surrogate in the new trial (optimized in the current trial and applied in the new trial) equals the average treatment effect on the final outcome in the new trial. Consequently, in a thought experiment where the current trial has 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.

In practice, an estimate of the optimal surrogate must be used as the actual surrogate endpoint. In Section 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 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 (e.g., van der Laan, Hubbard, and Pajouh, 2013), and is implemented in the SuperLearner R package. In Section 5 we further propose to update the super-learner fit of the optimal surrogate to solve an estimating equation [via targeted minimum loss-based estimation (TMLE)] that ensures 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. Whereas the TMLE update is advantageous compared to the untargeted super-learner estimator of the optimal surrogate given its asymptotic efficiency for the clinical parameter of interest θ_0, it 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 Web Appendix A 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 Theorem 2) 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; moreover this manuscript is about searching for a promising surrogate and does not address surrogate validation that is also of critical importance. In Section 6 we apply the proposed approach to two dengue vaccine efficacy trials. Web Appendix G studies the proposed approach in two simulations and Section 7 concludes with remarks.

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 (1989) replacement endpoint framework, the controlled direct and indirect causal effects framework (Robins and Greenland, 1992; Joffe and Greene, 2009), the principal stratification framework (Frangakis and Rubin, 2002), the meta-analysis framework (Daniels and Hughes, 1997; Buyse et al., 2000), and the causal selection diagram framework (Pearl and Bareinboim, 2011). 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 excellent 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. 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. Web Appendix B elaborates the connections of the optimal surrogate framework with the other surrogate frameworks.

The optimal surrogate approach also breaks new ground by searching for promising surrogates based on supervised nonparametric statistical learning. While historically pre-selected univariable or low-dimensional vector candidate surrogates are considered, the proposed approach allows all collected baseline and intermediate response data to potentially contribute to the optimal surrogate, selected and combined through unbiased machine learning, and not requiring parametric modeling assumptions.

2. Statistical Formulation of Estimation of an Optimal Surrogate

Let $O_i = (W_i, A_i, S_i, Y_i) \sim P_0$ for $i = 1,\ldots, n$ be the i.i.d. data, 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 is randomized conditional on W.

With S_a and Y_a potential outcomes under each treatment a, let $X = (W, S_0, S_1, Y_0, Y_1)$ denote the full-data structure, 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 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 future studies one hopes to replace the final outcome Y by a so-called surrogate outcome measured by the intermediate time point τ. At first we consider candidate surrogates as true unknown parameters, where we refer to any real-valued function $(W, A, S) \to \psi(W, A, S) \in \mathbb{R}$ as a candidate surrogate, representing a function of the true observed data generating distribution P_0 and of the random variables (W, A, S) collected 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. The key question is now how are we going to define a good surrogate, defined in terms of P_0? To start with we want the surrogate $S^{\psi} \equiv \psi(W, A, S)$ 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 $E_0\left(S_1^{\psi} - S_0^{\psi}\right) = 0$, where the counterfactual $S_a^{\psi} = \psi\left(W, a, S_a\right), a \in \left\{0, 1\right\}$. This guarantees that in this particular study involving sampling from P_0, a test for $H_0^{\psi}: E_0\left(S_1^{\psi} - S_0^{\psi}\right) = 0$, 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:

$$\psi \to MSE_{P_{X,0}}(\psi) \equiv \sum_a E_{P_{X,0}}\left\{g_0(a \middle| W)\left(Y_a - \psi\left(W, a, S_a\right)\right)^2\right\}. \quad (1)$$

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 participant is assigned treatment $A = a$ and one uses as surrogate outcome $S_a^{\psi} = \psi\left(W, a, S_a\right)$, 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, this particular weighting scheme $g_0(a \mid W)$ could be replaced by another weighting scheme. Given a class $\boldsymbol{\Psi}$ of possible surrogate functions $\psi()$, the P_0-optimal surrogate in this class is defined as

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

We focus on the nonparametric class $\boldsymbol{\Psi}$ consisting of all functions of (W, A, S). In this case, the choice of weight in $MSE_{P_{X,0}}\left(\mathrm{i.e.}, g_0(a \mid W)\right)$ 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

$$\psi_0^F(w, a, s) = E_0\left(Y_a \middle| W = w, S_a = s\right),$$

which is a standard solution to a minimization problem that is the same under and not under the Prentice definition constraint. The conditional randomization assumption implies 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, ψ_0^F is identifiable from P_0 and can also be defined as:

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

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 $E_{P0}(\psi_0(W, a, S_a) \mid W) = E_{P0}(Y_a \mid W)$, 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. 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:* $\bar{S}_{0,a} = E_0\left(Y_a \mid W, S_a\right), a \in \left\{0, 1\right\}$*, and* $E_{P_0}\left(\bar{S}_{0,a} \mid W\right) = E_{P_0}\left(Y_a \mid W\right)$.

This shows that the P_0-optimal surrogate has the perfect properties of a valid surrogate in the actual P_0-study. Moreover, if each treatment is considered separately, then the minimizer of $\psi_a \to MSE_{P_{X,a,0}}\left(\psi_a\right)$ 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 (1). Therefore the P_0-optimal surrogate is the same whether one minimizes the overall MSE in (1) or minimizes the treatment-specific MSEs separately (as we do in the application and simulations).

In practice, of course, the optimal surrogate cannot be used as a study endpoint, rather it must be estimated and the fitted values used. The statistical estimation problem for the original trial is now defined: we observe n i.i.d. $O \sim P_0 \in M$, the target parameter mapping is defined by $\Psi : M \to \Psi$ with $\Psi(P) = E_P(Y \mid W, A, S)$, and $\psi_0 = E_{P0}(Y \mid W, A, S)$ is the true value we aim to learn from the data.

3. Conditions on the New Study *P* Under Which the *P*$_0$-Optimal Surrogate is Also the *P*-Optimal Surrogate

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 $E_{P0}(Y \mid W, A, S)$ 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.

THEOREM 2: *Assume the current and future randomized studies defined above satisfy (1) Equal Conditional Means:* $E_P(Y^* \mid W^* = w, A^* = a, S^* = s) = E_{P_0}(Y \mid W = w, A = a, S = s)$ *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.*

Theorem 2 gives 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.

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

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 an important exception where Equal Conditional Means may more easily hold. The following theorem is merely a special case of the previous theorem, but its implication is that if the outcome Y only depends on the treatment through its effect on the surrogate vector S (i.e., Prentice's 'full mediation' criterion), then the new study can even consider a 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 the future treatment on the outcome, then the P_0-optimal surrogate can also be used in future studies evaluating different treatments, under a simpler Equal Conditional Means assumption that conditions on (W, S) but not on A.

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

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

Typically it is not reasonable to assume that the intermediate variable S completely blocks the effect of treatment (current and new) on the outcome, and even if it did, Equal Conditional Means may not hold. Web Appendix A discusses how $E_{P0}(Y| W, A, S)$ may still often be a good candidate surrogate for such a future study, and discusses implications about differences between $E_P(Y^*| W^* = w, A^* = a, S^* = s)$ and $E_{P_0}(Y \mid W = w, A = a, S = s)$.

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 $E_0(Y| W, A, S)$ with a minimizer of the risk of a loss: $\psi_0 = \text{argmin}_\psi P_0L(\psi)$, 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 (van der Laan, Polley, and Hubbard, 2007; van der Laan and Rose, 2011). We summarize how super-learner is used, with details provided in Web Appendix D. Super-learner operates by specifying a library of candidate estimators, and for each one computing the cross-validated risk (CV-RISK) [formula (1) in Web Appendix D] using squared error loss L(·) to be consistent with our proposed criterion (1) for the optimal surrogate. The discrete super-learner estimator is the candidate estimator with smallest CV-RISK and the super-learner is the convex combination of candidate estimators with smallest CV-RISK. Estimation of CV-RISK involves re-running the whole super-learner on learning samples and averaging estimates of the conditional risk on test samples.

One can also define a cross-validated R^2 (CV-R^2) taking values between 0 and 1 based on CV-RISK [formula (2) in Web Appendix D] that provides a universal measure of the strength of a given estimated surrogate $\widehat{\Psi}$, allowing us to compare different candidate surrogate estimators within and across studies. For example, one might construct a super-learner $\widehat{\Psi}_\delta$ based on δ-specific subsets (W_δ, A, S_δ) of the complete (W, A, S), where δ is a measure of the complexity of the resulting surrogate as a function of (W, A, S). One could now plot CV-R^2 of $\widehat{\Psi}_\delta$ 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 this surrogate in that trial; in practice some variable sets may be selected based on their high likelihood of being collected.

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 the super-learner estimator is advantageous 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.

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 $\epsilon_n = \arg\min_\epsilon P_n L\left(\psi_n^{\#}(\epsilon)\right)$ be the MLE, where P_n is the empirical distribution of the n observations and

$$L(\psi)(O) = -\{Y \log \psi(W, A, S) + (1 - Y)\log(1 - \psi(W, A, S))\} \quad (2)$$

is the log-likelihood loss function. This ϵ_n is easily calculated with a standard univariate logistic regression of Y on H_{g_n}, incorporating an o set. 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 predictive 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 (van der Laan and Rose, 2011). 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

$$0 = \frac{1}{n}\sum_{i=1}^{n} H_{g_n}\left(W_i, A_i\right)\left(Y_i - \psi_n^{\#}\left(W_i, A_i, S_i\right)\right), \quad (3)$$

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 the estimated optimal surrogate $\psi_n^{\#}$ is also a TMLE of the ATE on Y and is thus asymptotically linear and efficient for the ATE on Y.

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 $\left(W_i, A_i, \psi_n^{\#}\left(W_i, A_i, S_i\right)\right), i = 1, \ldots, n$, in our current study, compute the TMLE $\theta_{\psi_n^{\#}}^{TMLE}$ of the ATE $\theta_{\psi_n^{\#}} = \theta_{\psi_n^{\#}}^{1} - \theta_{\psi_n^{\#}}^{0} = E_0\left(\psi_n^{\#}\left(W, 1, S_1\right)\right) - E_0\left(\psi_n^{\#}\left(W, 0, S_0\right)\right)$. Under conditions, 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. Therefore, we wonder if this TMLE $\theta_{\psi_n^{\#}}^{TMLE}$ is also efficient for θ_0 based on observing $O = (W, A, S, Y)$? In other words, 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 $\theta_{\psi_n^{\#}}^{TMLE}$ of $\theta_{\psi_n^{\#}}$ and the TMLE $\tilde{\theta}_n^{TMLE}$ of θ_0 based on the full data (W, A, S, Y) including Y. From this it will be clear that $\theta_{\psi_n^{\#}}^{TMLE}$ 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 $\tilde{\theta}_n^{TMLE}$ of $E_0(Y_1 - Y_0)$ based on $O = (W, A, S, Y)$: 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) (in our setup with complete data on Y the efficient influence curve is the same with or without S). Let $\bar{Q}_n^0$ 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 (2), $\text{Logit}\,\bar{Q}_n^0(\epsilon) = \text{Logit}\,\bar{Q}_n^0 + \epsilon H_{g_n}$ be the least favorable submodel, and $\tilde{\epsilon}_n = \arg\min_\epsilon P_n L\left(\bar{Q}_n^0(\epsilon)\right)$ be the MLE of the fluctuation parameter ϵ. The TMLE of $\bar{Q}_0$ is defined as $\bar{Q}_n^1 = \bar{Q}_n^0(\tilde{\epsilon}_n)$ and the TMLE of the average treatment effect $E_0(Y_1 - Y_0)$ is given by $\tilde{\theta}_n^{TMLE} = \frac{1}{n}\sum_{i=1}^n \left\{\bar{Q}_n^1(W_i, 1) - \bar{Q}_n^1(W_i, 0)\right\}$. Due to the TMLE-update step we have that $\bar{Q}_n^1$ solves the score equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} H_{g_n}(W_i, A_i)\left(Y_i - \bar{Q}_n^1(W_i, A_i)\right), \quad (4)$$

and, as a result, the TMLE $\tilde{Q}_n^1 = \left(Q_{W,n}, \bar{Q}_n^1\right)$ (with $Q_{W,n}$ the empirical distribution of W) solves the efficient influence curve equation for $E_0(Y_1 - Y_0)$:

$$0 = \frac{1}{n}\sum_{i=1}^{n} D^{eff}\left(\tilde{Q}_n^1, g_n\right)(W_i, A_i, Y_i) = 0 \quad (5)$$

with $D^{eff}\left(\tilde{Q}_n^1, g_n\right)(W_i, A_i, Y_i) = D^{eff,1}\left(\tilde{Q}_n^1, g_n\right)(W_i, A_i, Y_i) - D^{eff,0}\left(\tilde{Q}_n^1, g_n\right)(W_i, A_i, Y_i)$, where $D^{eff,a}\left(\tilde{Q}_n^1, g_n\right)(W_i, A_i, Y_i) = \left(I(A_i = a)/g_n(a \mid W_i)\right)\left(Y_i - \bar{Q}_n^1(W_i, a)\right) + \bar{Q}_n^1(W_i, a) - \tilde{\theta}_n^{TMLE,a}$, and $\tilde{\theta}_n^{TMLE,a} = \frac{1}{n}\sum_{i=1}^n \bar{Q}_n^1(W_i, a)$ depends on both $\bar{Q}_n^1$ and $Q_{W,n}$. If we replace H_{gn} by a two dimensional $\bar{Q}_n^1$ with $H_{g_n}^a = I(A = a)/g_n(a \mid W)$, then the updated $\bar{Q}_n^1 = \bar{Q}_n^0(\tilde{\epsilon}_n)$ (where $\tilde{\epsilon}_n$ is now a two dimensional parameter) also yields a TMLE for the bivariate parameter (EY_0, EY_1) (the above TMLE targets the difference), which solves $0 = P_n D^{eff,a}\left(\tilde{Q}_n^1, g_n\right)$ for each $a = 0, 1$. We use such treatment-specific TMLEs because in the application we estimate non-additive difference treatment effects (i.e., relative risk EY_1/EY_0). These equations are standard TMLE equations (e.g., defined in van der Laan and Rose, 2011, p. 527–529), and are the basis for the double robustness and asymptotic efficiency of the TMLEs.

TMLE $\theta_{\psi_n^\#}^{TMLE}$ of the ATE $\theta_{\psi_n^\#}$ on the estimated optimal surrogate $\psi_n^\#$ based on $O^r = \left(W, A, \psi_n^\#(W, A, S)\right)$: 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 initial estimator of $\bar{Q}_0(W, A) = E_0\left(\psi_0(W, A, S) \mid W, A\right) = E_0(Y \mid W, A)$, where one might again use super-learning. Let us denote this estimator with $\bar{Q}_n^{\#0}$. This is nothing else than an estimator of $\bar{Q}_0(W, A) = E_0\left(E_0(Y \mid W, A, S) \mid W, A\right)$, 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 $\text{Logit}\, \bar{Q}_n^{\#0}(\epsilon) = \text{Logit}\, \bar{Q}_n^{\#0} + \epsilon H_{g_n}$, and defines $\epsilon_{n1} = \arg\min_{\epsilon} \sum_{i=1}^{n} L_1\left(\bar{Q}_n^{\#0}(\epsilon)\right)\left(O_i^r\right)$ solves the following score equation (analog to (4)):

$$0 = \frac{1}{n} \sum_{i=1}^{n} H_{g_n}\left(W_i, A_i\right)\left(\psi_n^{\#}\left(W_i, A_i, S_i\right) - \bar{Q}_n^{\#1}\left(W_i, A_i\right)\right). \quad (6)$$

The TMLE $\theta_{\psi_n^{\#}}^{TMLE}$ of $\theta_{\psi_n^{\#}}$ is now the substitution estimator

$$\theta_{\psi_n^{\#}}^{TMLE} = \frac{1}{n} \sum_{i=1}^{n} \left\{\bar{Q}_n^{\#1}\left(W_i, 1\right) - \bar{Q}_n^{\#1}\left(W_i, 0\right)\right\}.$$

Now we utilize the fact that $\psi_n^{\#}$ was targeted so that it solves the equation (3). Equation (3) combined with the score equation (6) implies that $\bar{Q}_n^{\#1}$ solves

$$0 = \frac{1}{n} \sum_{i=1}^{n} H_{g_n}\left(W_i, A_i\right)\left(Y_i - \bar{Q}_n^{\#1}\left(W_i, A_i\right)\right). \quad (7)$$

Thus, this TMLE $Q_n^1 = \left(Q_{W,n}, \bar{Q}_n^{\#1}\right)$ also solves the efficient influence curve equation for θ_0:

$$0 = \frac{1}{n} \sum_{i=1}^{n} D^{eff}\left(Q_n^1, g_n\right)\left(W_i, A_i, \psi_n^{\#}\left(W_i, A_i, S_i\right)\right) = 0. \quad (8)$$

(And parallel to the above, the TMLE $\theta_{\psi_n^{\#}}^{TMLE,a}$ of $\theta_{\psi_n^{\#}}^{a}$ solves $0 = P_n D^{eff,a}\left(Q_n^1, g_n\right)$ with $D^{eff,a}\left(Q_n^1, g_n\right)\left(W_i, A_i, \psi_n^{\#}\left(W_i, A_i, S_i\right)\right) = \left(I\left(A_i = a\right)/g_n(a \mid W_i)\right)\left(Y_i - \bar{Q}_n^{\#1}\left(W_i, a\right)\right) + \bar{Q}_n^{\#1}\left(W_i, a\right)$ Thus, $-\theta_{\psi_n^{\#}}^{TMLE,a}$.)

$\theta_{\psi_n^{\#}}^{TMLE}$ 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 $\bar{Q}_n^{\#0}$ of $\bar{Q}_0$ (as discussed above, involving first regressing Y on W, A, S and then regressing that fit on W, A). 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 M.

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

Thus, 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 $\theta_0 = E_0(Y_1 - Y_0)$ based on the observed data (W, A, S, 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 (and for other contrasts of EY_0 and EY_1). 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 achieve 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 ATE on our targeted surrogate based in a new study would not result in an asymptotically efficient estimator of the ATE on Y^*. Nevertheless, it is an appealing property of the estimated optimal surrogate that in the current study it yields an asymptotically efficient estimator of the average clinical treatment effect.

6. Application to Two 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 (Capeding et al., 2014) and Latin America (Villar et al., 2015), 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 participants with Month 13 titer data, which were measured in a random sample of study participants and in all participants 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 participant'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, where in CYD15 we only include data from 9–14 year-olds to increase the credibility of the contained support assumption of Theorem 2.

We first calculate 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 $\theta_{\psi_n^{\#}}^{a} = E_0\left(\psi_n^{\#}\left(W, a, S_a\right)\right)$ 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, $VE_{\psi_n^{\#}} = 1 - \theta_{\psi_n^{\#}}^{1}/\theta_{\psi_n^{\#}}^{0}$. Wald 95% confidence intervals for each $\theta_{\psi_n^{\#}}^{a}$ are calculated by estimating the variance of each $\theta_{\psi_n^{\#}}^{TMLE,a}$ by the sample variance of the efficient influence curve values $D^{eff,a}\left(Q_n^1, g_n\right)\left(W_i, A_i, \psi_n^{\#}\left(W_i, A_i, S_i\right)\right)$ defined above. The delta method is then applied to obtain the variance of $\log\left(\theta_{\psi_n^{\#}}^{TMLE,1}/\theta_{\psi_n^{\#}}^{TMLE,0}\right)$ and the resulting symmetric Wald 95% confidence limits are transformed to obtain the CI for $VE_{\psi_n^{\#}}$. The same approach to obtain Wald CIs is used for $E_0(Y_0)$, $E_0(Y_1)$, and $\theta_0 = 1 - E_0(Y_1)/E_0(Y_0)$ based on (W_i, A_i, Y_i), with values $D^{eff,a}\left(Q_n^1, g_n\right)\left(W_i, A_i, \psi_n^{\#}\left(W_i^*, A_i^*, S_i^*\right)\right)$ replaced with $D^{eff,a}\left(\tilde{Q}_n^1, g_n\right)\left(W_i, A_i, Y_i\right)$.

Second, we calculate the $\psi_n^{\#}\left(W_i^*, A_i^*, S_i^*\right)$ surrogate outcome values for the n^* CYD15 participants (with $\psi_n^{\#}(\cdot)$ calculated from CYD14), and, based on the CYD15 data $\left(W_i^*, A_i^*, S_i^* \psi_n^{\#}\left(W_i^*, A_i^*, S_i^*\right)\right)$, estimate the treatment-specific surrogate means in CYD15: $\theta_{\psi_n}^a(P) = E_P\left(E_P\left(\psi_n^{\#}(W^*, a, S^*) \mid W^*, A^* = a\right)\right)$ for a = 0,1 and $VE_{\psi_n^{\#}}(P) = 1 - \theta_{\psi_n^{\#}}^1(P)/\theta_{\psi_n^{\#}}^0(P)$. Here the TMLE $\theta_{\psi_n}^{TMLE,a}(P)$ of $\theta_{\psi_n^{\#}}^a(P)$ is the solution to $0 = P_{n^*} D^{eff,a}\left(Q_n^1, g_n\right)$. Lastly, to check how well the estimated optimal surrogate performs in its use to estimate the clinical parameters in the new trial, we compare the TMLEs of the surrogate parameters to the TMLEs of $E_P\left(Y_0^*\right), E_P\left(Y_1^*\right)$, and $\theta_P^* = VE_P^* = 1 - E_P\left(Y_1^*\right)/E_P\left(Y_0^*\right)$, calculated based on the CYD15 data $\left(W_i^*, A_i^*, Y_i^*\right)$, where $\tilde{\theta}_{n^*}^{TMLE,a}(P)$ is the TMLE of $E_P\left(Y_a^*\right)$. Wald 95% confidence intervals for the $E_P\left(Y_a^*\right)$ and $VE^*(P)$ parameters based on $\left(W_i^*, A_i^*, Y_i^*\right)$ are computed in the identical way as done for CYD14. The CIs for the surrogate parameters in CYD15 are computed similarly, where the variance of each $\theta_{\psi_n^{\#}}^{TMLE,a}(P)$ for $a \in \{0, 1\}$ is estimated by the sample variance of the n^* values $D^{eff,a}\left(Q_n^1, g_n\right)\left(W_i^*, A_i^*, \psi_n^{\#}\left(W_i^*, A_i^*, S_i^*\right)\right)$.

6.1 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. Table 1 displays the input variables, learner types, and pre-screening approaches applied to each learner type for estimating $\psi_0 = E_0(Y \mid W, A = a, S)$. Figure 1 shows point and 95% CI estimates of the cross-validated MSEs (van der Laan, Hubbard, and Pajouh, 2013) for each individual statistical algorithm as well as for discrete super-learner and super-learner. A logistic regression model (glm) after variable screening that disallows PRNT$_{50}$ titers performs best (with the lowest CV-MSE) for each treatment group (Table 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 CV-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 from CYD14 data as described in Section 5. Figure 2(a) shows empirical reverse cdf plots of $\psi_n^{\#}\left(W_i, A_i = a, S_i\right)$ 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 $\psi_n^{\#}(W, A, S)$,

$\hat{E}_0\left(Y_1\right) = \theta_{\psi_n^{\#}}^{TMLE,1} = 0.017$ (95% CI 0.016 - 0.019),

$\hat{E}_0(Y_0) = \theta_{\psi_n^\#}^{TMLE,0} = 0.039$ (95% CI $0.036 - 0.042$), and $\widehat{VE}_0 = \theta_{\psi_n^\#}^{TMLE} = 55\%$ (95% CI $49 - 61$).

These estimates are close to those obtained based on (W_i, A_i, Y_i), with $\tilde{\theta}_n^{TMLE,1} = 0.017$ (95% CI $0.014 - 0.021$), $\tilde{\theta}_n^{TMLE,0} = 0.039$ (95% CI $0.031 - 0.047$), and $\widehat{VE}_0 = \tilde{\theta}_n^{TMLE} = 55\%$ (95% CI $40 - 66$) as they should be based on the results in Section 5.

6.2 Applying the estimated optimal surrogate from the original trial to the new trial

Figure 2(b) shows empirical reverse cdf plots of $\psi_n^\#(W_i^*, A_i^* = a, S_i^*)$ for each treatment $a \in \{0, 1\}$ by case-control status $y \in \{0, 1\}$ in CYD15, showing diminution of classification accuracy of the estimated optimal surrogate built on CYD14 for the new study CYD15 (as expected). Table 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 $VE_{\psi_n^\#}^{TMLE}(P) = 66\%$ (95% CI $58 - 72$) and $\widehat{VE}_P^* = \tilde{\theta}_{n*}^{TMLE}(P) = 61\%$ (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 stems from imperfect adherence to the Theorem 2 assumptions. The diagnostic analysis in Web Appendices E–F supports that the assumptions were approximately satisfied, with only minor violations, which was made possible by the fact that CYD14 and CYD15 were essentially the same protocol implemented in two geographic regions.

7. Discussion

VanderWeele (2013) and discussants Joffe (2013) and Pearl (2013) 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 (2013) shows that commonly used methods for surrogate endpoint evaluation generally do not guarantee avoiding this paradox. The first useful feature of the newly proposed approach is that it starts at this minimal requirement, defining the optimal surrogate in a way guaranteed to satisfy the Prentice definition of a valid surrogate within the original trial and thus avoid the paradox (and then the estimated optimal surrogate (EOS), which can be used as a surrogate endpoint in practice, satisfies the Prentice definition in large samples). As such the proposed approach responds to Pearl's (2013) 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?…" A second useful feature of the approach is that the treatment effect on the EOS has the same interpretation as the treatment effect on the clinical endpoint of interest.

A third useful feature of the proposed approach is that the EOS– in being built by super-learner followed by a TMLE update– contains all information about the average clinical treatment effect in the original trial. A fourth useful feature is the approach's use of super-learner with its principled cross-validation approach to build and compare best models for estimating the optimal surrogate. Super-learner is useful for applications where multiple baseline covariates and/or intermediate response endpoints are measured, yet there is considerable 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.

A challenge posed to the framework is that through super-learner the EOS may be based on a complicated combination of models that is hard to interpret. This underscores the importance of building multiple EOSs from different input variable sets ranging from single-variable to all-variable models, where cross-validation criteria allow principled selection of a most parsimonious EOS with near-optimal predictive performance. A related challenge is that researchers in future trials may not have access to the code used by the previous researchers to calculate the EOS. This may require use of an open research paradigm where web calculators are made available that input (W, A, S) values and output EOS values.

This article considers an ideal setting with no missing data and where the clinical outcome is never observed before the intermediate response endpoints are measured. Moreover, we used a particular loss function for defining optimal prediction. Future work is of interest to accommodate these issues. Theorems 2 and 3 provide conditions for using the EOS from an original trial to confer correct estimation of the clinical treatment effect in a new setting/trial based on this surrogate endpoint without measuring the clinical endpoint. The inference part of these results hold for an infinite original trial, such that 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 EOS is modeled parametrically but not if modeled nonparametrically. Importantly, because in many practical applications the critical assumption of our Theorems 2 and 3 for making valid inferences for a new setting– Equal Conditional Means– is implausible or dubious, a utility of the theorems is in clarifying why direct clinical endpoint studies are generally needed. Additional research is of interest to allow deviations from the theorem assumptions. Moreover, additional research may consider applications where a set of randomized clinical efficacy trials are available that provide direct clinical endpoint data for estimating how the conditional means vary over settings, which could allow new transportability results under weaker assumptions. Dummy versions of the dengue application data sets and R code producing all of the (dummy) data results is provided in Web Appendix H.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank the participants of the CYD14 and CYD15 trials and our SanofiPasteur colleagues who conducted these trials.

References

Buyse M, Molenberghs G, Burzykowski T, Renard D, and Geys H (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 1, 49–67. [PubMed: 12933525]

Capeding MR, Tran NH, Hadinegoro SRS, Ismail HIHM, Chotpitayasunondh T, Chua MN, et al. (2014). 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, 1358–1365.

Daniels M and Hughes M (1997). Meta-analysis for the evaluation of potential surrogate markers. Statistics in Medicine 16, 1965–1982. [PubMed: 9304767]

Frangakis C and Rubin D (2002). Principal stratification in causal inference. Biometrics 58, 21–29. [PubMed: 11890317]

Joffe M (2013). Discusion on “surrogate measures and consistent surrogates”. Biometrics 69, 572–575. [PubMed: 24073863]

Joffe M and Greene T (2009). Related causal frameworks for surrogate outcomes. Biometrics 65, 530–538. [PubMed: 18759836]

Pearl J (2013). Discussion on “surrogate measures and consistent surrogates”. Biometrics 69, 573–577.

Pearl J and Bareinboim E (2011). 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.

Prentice R (1989). Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in Medicine 8, 431–440. [PubMed: 2727467]

Robins J and Greenland S (1992). Identifiability and exchangeability of direct and indirect effects. Epidemiology 3, 143–155. [PubMed: 1576220]

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

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

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

Villar L, Dayan GH, Arredondo-García JL, Rivera DM, Cunha R, Deseda C, et al. (2015). Efficacy of a tetravalent dengue vaccine in children in Latin America. New England Journal of Medicine 372, 113–123. [PubMed: 25365753]

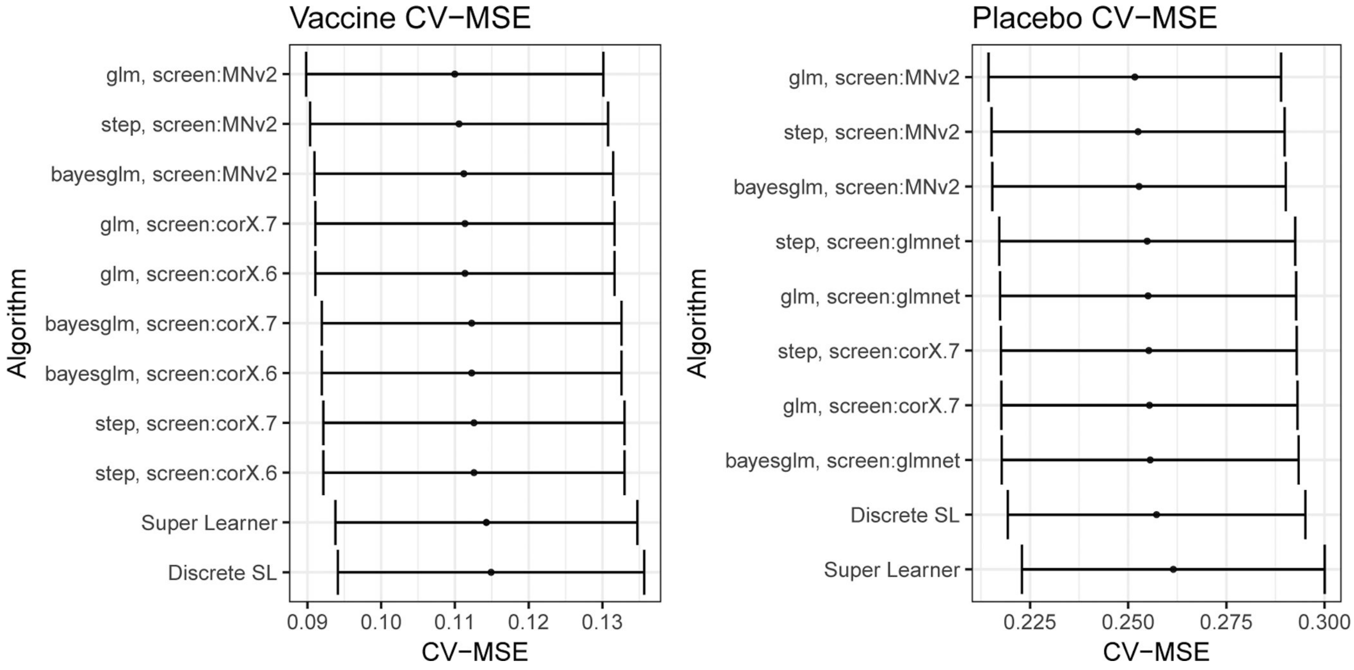

Figure 1.

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

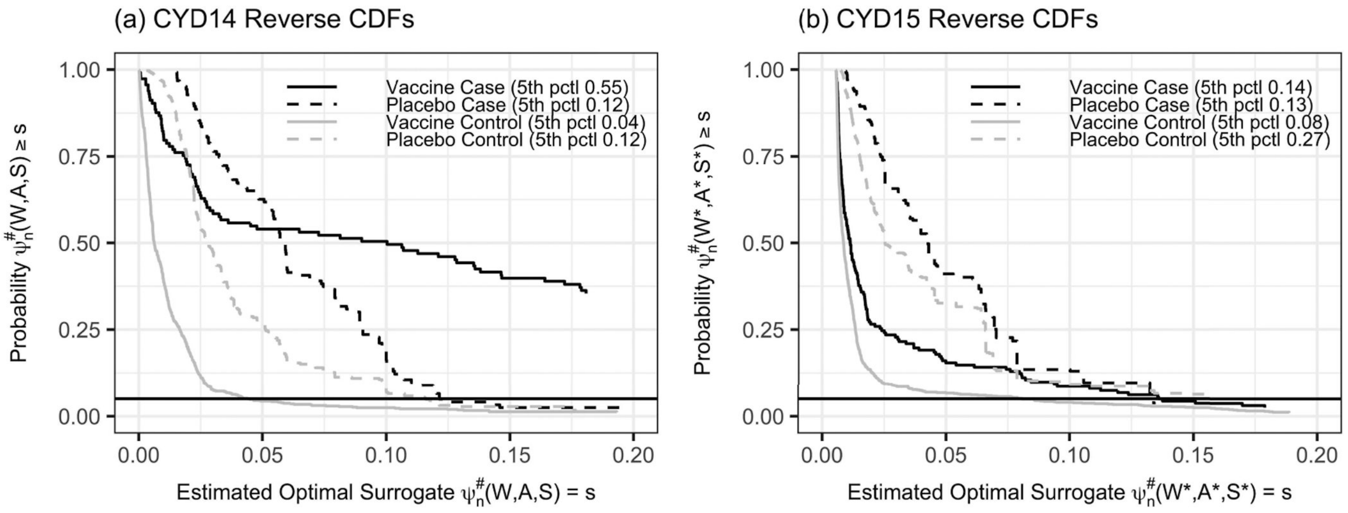

Figure 2.

(a) Empirical reverse cumulative distribution functions (cdfs) of the estimated optimal surrogate $\psi_n^{\#}\left(W_i, A_i = a, S_i\right)$ 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 $\psi_n^{\#}\left(W_i, A_i = a, S_i^*\right)$ 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 (not surprisingly).

Table 1

Input variables, screens, and learner types used in the super-learner for the CYD14 dengue vaccine efficacy trial (35 total statistical algorithms for estimating $\psi_0 = E_0(Y|W, A, S)$ defined by screens crossed with learner types).

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 Version 2 (V2) 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` (i.e., lasso)
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 V2 titer variables
screen.MNv2	Disallow PRNT_{50} titer variables
Learner Types	Boldfaced courier-font learning algorithms (e.g., `SL.mean`) are available in the SuperLearner R package available at CRAN
`SL.mean`	$E_0(Y \mid W, A = a, S)$[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`.

Table 2

Best performing models for estimating $\psi_0 = E_0(Y \mid W, A, S)$ 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 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 estimated serotype frequency of a participant’s country.

Table 3

Comparison of inferences on the surrogate parameters $\theta^{a}_{\psi_n^{\#}}(P) \equiv E_P\left(E_P\left(\psi_n^{\#}(W^*, a, S^*) \mid W^*, A^* = a\right)\right)$ for each a ∈ {0,1} and $VE_{\psi_n^{\#}}(P) = 1 - \theta^{1}_{\psi_n^{\#}}(P)/\theta^{0}_{\psi_n^{\#}}(P)$ based on $\left(W^*, A^*, \psi_n^{\#}(W^*, A^*, S^*)\right)$ versus direct inferences on the clinical dengue endpoint parameters $E_P\left(Y_a^*\right)$ and $\theta_P^* = VE_P^* = 1 - E_P\left(Y_1^*\right)/E_P\left(Y_0^*\right)$ in CYD15. Included is a summary of enrollment numbers, incidence of VCD, and number of participants 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)	$E_P\left(Y_1^*\right)$	0.014 (95% CI 0.012–0.017)
$\theta^{0}_{\psi_n^{\#}}(P)$	0.057 (95% CI 0.049–0.065)	$E_P\left(Y_0^*\right)$	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 $\tilde{\theta}^{TMLE,1}_{n*}(P)$, $\tilde{\theta}^{TMLE,0}_{n*}(P)$, and $\widetilde{VE}_{n*}(P) = 1 - \tilde{\theta}^{TMLE,1}_{n*}(P)/\tilde{\theta}^{TMLE,0}_{n*}(P)$.

[c] Measured in 98.3% and 99.8% of endpoint cases with $Y = 1$ or $Y^* = 1$ for CYD14 and CYD15, respectively.