Biostatistics (2024), **0**, 0, *pp.* 1–**??**
doi:10.1093/biostatistics/output

Supplementary Material For "A Surrogate Endpoint Based Provisional Approval Causal Roadmap"

Peter B. Gilbert [1,2,*], James Peng [2], Larry Han [3], Theis Lange [4], Yun Lu [5], Lei Nie [6], Mei-Chiung Shih [7], Salina P. Waddy [8], Ken Wiley [8], Margot Yann [9], Zafar Zafari [10,11], Debashis Ghosh [12], Dean Follmann [13], Michal Juraska [1], Iván Díaz [14], for the Forum on the Integration of Observational and Randomized Data

[1] *Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Center, Seattle, WA, 98109, USA*

[2] *Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA*

[3] *Department of Public Health and Health Sciences, Bouvé College of Health Sciences, Northeastern University, MA, 02115, USA*

[4] *Department of Public Health, University of Copenhagen, Copenhagen, K-1353, Denmark*

[5] *Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, 20993, USA*

[6] *OB/OTS/CDER, Food and Drug Administration, Silver Spring, MD, 20993, USA*

[7] *VA Palo Alto Health Care System, 795 Willow Road, Menlo Park, CA, 94025, USA*

[8] *Division of Clinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, Maryland, 20892, USA*

[9] *School of Public Health, UC Berkeley, Forum for Collaborative Research, UCDC Campus, Washington, DC, 20036, USA*

[10] *University of Maryland School of Pharmacy, Baltimore, MD, 21201, USA*

[11] *University of Maryland Institute for Health Computing, Bethesda, MD, 20852, USA*

[12] *Department of Biostatistics & Informatics, Colorado School of Public Health, Aurora, CO, 80045, USA*

[13] *Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Rockville, MD, 20852, USA*

[14] *Department of Biostatistics, New York University Grossman School of Medicine, New York, NY, 10016, USA*

A. Comparison of the present work to Athey et al. (2024)

We detail differences between the present manuscript and Athey *and others* (2024) in their providing general statistical methodology for the same or similar transportability objective.

1. The two articles use similar notation except our study indicator Z is their $\mathcal{P}$ and our binary treatment A is their W. In the following notes we use the notation of the present article.

2. The two articles consider the same target causal parameter of interest – TE – a contrast in $E[Y(1)|Z = 0]$ vs. $E[Y(0)|Z = 0]$. This work develops all elements allowing for a general contrast function $h(x, y)$ satisfying $h(x, y) = 0$ if and only if $x = y$, given that a multiplicative contrast such as $h(x, y) = \log(E[Y(1)|Z = 0], E[Y(0)|Z = 0])$ or $h(x, y) = 1 - E[Y(1)|Z = 0]/E[Y(0)|Z = 0]$ is needed for our provisional approval application, whereas Athey et al. restrict attention to an additive difference contrast $h(x, y) = E[Y(1)|Z = 0] - E[Y(0)|Z = 0]$. The Athey et al. results could be readily generalized to handle a general contrast.

3. This work supposes all observational study participants have treatment level $A = 0$ known whereas Athey et al. assume A is missing/unknown. An implication is Athey et al.'s Assumption 1 of a single random sample is different from our set-up. Under both the Comparability Assumption and the Surrogacy Assumption (perfect surrogate version) noted below, this difference does not affect the identifiability results nor the estimators, such that results of the two articles are equivalent in this case. Under violations of either of these assumptions (the focus for our provisional approval objective), the identifiability results and hence also the estimators are different for the two articles.

4. Related to the previous point, both articles provide a nonparametric efficient influence function for use in estimation, which are equal under both the Comparability Assumption and Surrogacy Assumption (prefect version) and differ otherwise due to the different set-up.

5. Both articles use the "optimal surrogate" (our nomenclature from Price *and others* (2018)) or equivalently "surrogate index" (Athey et al. nomenclature) as a central ingredient of the results: $E[Y|X, Z = 0, A = a, S]$. Indeed, a significant idea common to both articles is to make use of a conditional regression that can depend on multivariable S, an idea that we previously championed in Price *and others* (2018) and picked up in this work. Athey et al. provide a valuable summary of the benefits of multiple surrogates in their Section 3.2.3.

6. Both articles assume that in the experimental study $Z = 0$, treatment assignment A is unconfounded (strong ignorability), equivalently the $Z = 0$ study is randomized within levels of baseline characteristics X. This assumption is A2 in this work and Assumption 2 in Athey et al.

7. Both articles make a Comparability Assumption, in different ways. Our assumption A4 equates two conditional means offset by the Untreated-to-Control transport bias function $u^{UC}(X, S)$, and Athey et al.'s Assumption 4 expresses conditional independence.

8. Both articles use a Surrogacy Assumption (a Prentice valid surrogate: our assumption A6; Athey et al. Assumption 3) as a key assumption for the results, with difference that our assumption is expressed in terms of equal conditional means and the Athey et al. assumption is expressed as conditional independence. For identifiability of the target causal parameter of interest, the assumption of equal conditional means suffices, with conditional independence not necessary; we speculate that it was not Athey et al.'s goal to define minimally sufficient identifiability assumptions as they had other reasons to prefer to use the stronger conditional independence assumption. In addition, the present article does not state the Surrogacy Assumption apart from a bias function given that for the provisional approval application the data analysis prioritizes scenarios with a non-zero bias function.

9. Elaborating on the last point, the two articles make a different "style choice" regarding how

to include bias functions in relation to the key Surrogate Assumption and Comparability Assumption. Athey et al. introduces these key assumptions under the ideal case that both assumptions hold, and provide results expressing the bias that results from deviations from these assumptions. The present work includes the bias functions as fixed and user-specified functions directly in the identifiability assumptions. Our provisional approval application drove this choice, where for this application data analysis assuming bias is most germane based on our understanding that regulators will generally require explicit conservative inference to provide sufficient evidence undergirding a provisional approval decision, where the pre-specified success benchmark in the statistical analysis plan would include non-zero bias functions.

10. Athey et al. included three representations of the statistical estimand of interest [their equation (4.1) and our equation (3.4)] using a surrogate index, a surrogate score [their Definition 2: surrogate score $P(A = 1|S = s, X = x, Z = 0)$], or both, with utility that each representation indicates different parameters that require estimation and for some applications one or another representation may be more advantageous. These representations are equations (4.2), (4.3), (4.4) in Athey et al. Our article focuses on a representation most similar to (4.2), and in the special case that the Comparability Assumption and the Surrogacy Assumption (perfect version) hold, it is straightforward to write our statistical estimand in any of the three representations expressed by Athey et al.

11. This work emphasizes the fact that the surrogates S are measured via a two-phase sampling design in both studies, developing all results accounting for this ubiquitous data reality for the provisional approval application. Athey et al. considers complete data on the surrogates S in the observational study; the results of Athey et al. could be extended to account for two-phase sampling of S under a missing at random assumption, where there are multiple ways that the extensions could account for the missing data on S.

12. This work considered the issues of the target outcome $Y = I(T \leqslant t_0)$ being subject to right-censoring and considered Intercurrent events (ICEs), complications not considered by Athey et al.

B. Violation of the Simplifying Assumption that All Enrolled Participants in Both Studies were Free of the Target Outcome through the Visit at τ for Surrogate Measurement

Our development made the simplifying assumption that all enrolled participants in both studies did not experience the target outcome by the visit τ at which intermediate outcomes S are measurable, i.e., $Y_i^{0-\tau} = 0$ for all i. In practice this is likely violated. We also made the simplifying assumption of no loss to follow-up before τ. To address this second issue, the methods are valid when applied only including participants observed to reach time τ, by adding a random censoring assumption conditional on (X, Z, A).

To handle early target outcome events Y before τ, for applications where the surrogate is only meaningfully defined for participants reaching τ free of the target outcome, principal stratification may be an appropriate framework for extensions, as noted in Section 3.4.4, in which case the causal parameters would be defined for the always-survivors (AS) principal stratum with $Y^{0-\tau}(1) = Y^{0-\tau}(0) = 0$. Under the "equal early clinical risk" (EECR) assumption of no individual-level causal treatment effects on Y by τ $[P(Y^{0-\tau}(1) = Y^{0-\tau}(0)) = 1]$, then the approach to data analysis that simply excludes all early failure events works, as in this case both causal parameters $E[Y(a)|Z = 0, Y^{0-\tau}(a) = 0]$ equal $E[Y(a)|Z = 0, Y^{0-\tau}(1) = Y^{0-\tau}(0) = 0]$, for $a = 0, 1$.

For applications where it is not appropriate to assume EECR, an alternative assumption that achieves the same objective as EECR is that both of the following conditions hold with probability one:

1. $Y(1)$ and $Y^{0-\tau}(0)$ are independent conditional on $(X, S(1))$, $Y^{0-\tau}(1) = 0$ and $A = 1$;

2. $Y(0)$ and $Y^{0-\tau}(1)$ are independent conditional on $(X, S(0))$, $Y^{0-\tau}(0) = 0$ and $A = 0$,

where recall the time origin for $Y(a) := I(T(a) \leqslant t_0)$ is τ and $Y^{0-\tau}(a)$ is the indicator of failure after enrollment by τ under assignment a, for $a = 0, 1$. This assumption implies that conditioning on $Y^{0-\tau}(a) = 0$ and X is the same thing as conditioning on $Y^{0-\tau}(1) = Y^{0-\tau}(0) = 0$ and X, as noted in Shepherd et al. (2006). For arm $A = 1$, it requires that the risk of the target outcome under assignment $A = 1$ is the same in the $\{Y^{0-\tau}(1) = Y^{0-\tau}(0) = 0\}$ AS principal stratum as in the $\{Y^{0-\tau}(1) = 0, Y^{0-\tau}(0) = 1\}$ principal stratum within levels of X and $S(1)$ (for participants with $Y^{0-\tau} = 0$). A similar interpretation applies for arm $A = 0$. This assumption (e.g., for $A = 1$) can be violated if experiencing the target outcome by τ under control assignment correlates with experiencing the target outcome by τ under treatment assignment, which could occur due to exposure or biological susceptibility factors not fully captured in the baseline covariates X. Going beyond the assumptions considered above would require additional sensitivity analysis that would further enlarge the estimated uncertainty intervals about treatment efficacy.

C. Mapping of causal parameters to statistical estimands (identifiability)

To show identification of $E[Y(0)|Z = 0]$ in equation (3.4), we consider the parameter within levels of baseline covariates X:

$$
\begin{aligned}
E[Y(0)|X, Z = 0] &= E[Y(0)|X, Z = 0, A = 0] && \text{(randomization A2)} \\
&= E[E[Y(0)|X, Z = 0, A = 0, S(0)]|X, Z = 0, A = 0] && \text{(iterative expectation)} \\
&= E[E[Y(0)|X, Z = 1, A = 0, S(0)]|X, Z = 0, A = 0] \\
&\qquad - E[\mu^{UC}(X, S(0))|X, Z = 0, A = 0] && \text{(A4, A5)} \\
&= E[g(X, S)|X, Z = 0, A = 0] \\
&\qquad - E[\mu^{UC}(X, S)|X, Z = 0, A = 0]. && \text{(causal consistency A1)}
\end{aligned}
$$

By averaging $E[Y(0)|X, Z = 0]$ with respect to the distribution of X conditional on $Z = 0$, we

then obtain:

$$E[Y(0)|Z=0] = E\big\{E[g(X,S)|X,Z=0,A=0] \mid Z=0\big\}$$
$$- E\big\{E[\mu^{UC}(X,S)|X,Z=0,A=0] \mid Z=0\big\}.$$

Lastly, from Rose and van der Laan (2011), $g(X,S) := E(Y \mid X, Z=1, A=0, S)$ and each $g_a^*(X,S)$ defined in Section 3.4 is identified by equation (3.5).

For identification of $E[Y(1)|Z=0]$ in equation (3.6), let $H^{01}(s|x)$ be the cdf of $S(1)$ conditional on $X, Z=0, A=1$. Calculations show:

$$E[Y(1)|X=x,Z=0] = E[Y(1)|X=x,Z=0,A=1] \quad \text{(A2)}$$
$$= E[E[Y(1)|X=x,Z=0,A=1,S(1)]|X=x,Z=0,A=1] \quad \text{(iter. exp.)}$$
$$= \int E[Y(1)|X=x,Z=0,A=1,S(1)=s]dH^{01}(s|x)$$
$$= \int \big\{E[Y(0)|X=x,Z=0,A=0,S(0)=s] + u^{CT}(x,s)\big\}\, dH^{01}(s|x) \quad \text{(A2, A3, A6)}$$
$$= \int \big\{E[Y(0)|X=x,Z=1,A=0,S(0)=s] + u^{CT}(x,s) - u^{UC}(x,s)\big\}\, dH^{01}(s|x) \quad \text{(A4, A5)}$$
$$= E[E[Y(0)|X=x,Z=1,A=0,S(0)]|X=x,Z=0,A=1]$$
$$+ E[\mu^{CT}(x,S)|X=x,Z=0,A=1] - E[\mu^{UC}(x,S)|X=x,Z=0,A=1]$$
$$= E[g(x,Z=1,S)|X=x,Z=0,A=1]$$
$$+ E[\mu^{CT}(x,S)|X=x,Z=0,A=1] - E[\mu^{UC}(x,S)|X=x,Z=0,A=1]. \quad \text{(A1)}$$

The same as done for $E[Y(0)|Z=0]$, by averaging $E[Y(1)|X,Z=0]$ with respect to the distribution of X conditional on $Z=0$, we obtain:

$$E[Y(1)|Z=0] = E\big\{E[g(X,S)|X,Z=0,A=1] \mid Z=0\big\}$$
$$+ E\big\{E[\mu^{CT}(X,S) - \mu^{UC}(X,S)|X,Z=0,A=1] \mid Z=0\big\}.$$

Lastly, exactly as for $E[Y(0)|Z=0]$, from Rose and van der Laan (2011), $g(X,S) := E(Y \mid X, Z=1, A=0, S)$ and each $g_a^*(X,S)$ defined in Section 3.4 is identified by equation (3.5).

D. Sandwich variance estimation for the plug-in estimator

In this section, we describe how to calculate the sandwich variance for the plug-in estimators (Stefanski and Boos, 2002). For simplicity, we consider the no-bias function scenario, although our R code allows for the bias functions. Our plug-in estimators are defined as:

$$\widehat{\theta}_{a,\text{plug-in}} = \frac{1}{n_{RCT}} \sum_{i=1}^{n} I(Z_i = 0) \widehat{E}[\widehat{g}(X,S)|X_i, Z_i = 0, A_i = a].$$

To ease notation, we write $g(x,s) := g(X=x, Z=1, S=s)$ and define $\mu_a(x) := E[g(X,S)|X = x, Z=0, A=a]$, for $a = 0, 1$. Further, from previously defined notation, we have sampling probabilities $\pi(x,z,a,t,\delta) := P(\epsilon_s = 1|X=x, Z=z, A=a, \tilde{T}=t, \Delta=\delta)$.

Suppose that we have parametric models for nuisance functions g, μ_a, and $\pi(\cdot)$, with parameters β, γ_a, and α, respectively. We therefore write our set of nuisance functions as $g(x,s;\beta)$, $\mu_0(x;\gamma_0)$, $\mu_1(x;\gamma_1)$, and $\pi(x,z,a,t,\delta;\alpha)$. Let $\hat{\beta}$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\alpha}$ denote estimators for these parameters obtained through estimating equations. First, let h_π denote the estimating function used to obtain $\hat{\alpha}$:

$$0 = \sum_{i=1}^{n} h_\pi(X_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \alpha).$$

Next, h_g^C, $h_{\mu_0}^C$, and $h_{\mu_1}^C$ denote the estimating functions used to obtain $\hat{\beta}$, $\hat{\gamma}_0$, and $\hat{\gamma}_1$ assuming complete data. To account for incomplete sampling of S, we use inverse probability sampling (IPS) weighting, which gives us our final estimating equations:

$$0 = \sum_{i=1}^{n} h_g(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \beta, \alpha)$$

$$0 = \sum_{i=1}^{n} h_{\mu_0}(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \alpha, \beta, \gamma_0)$$

$$0 = \sum_{i=1}^{n} h_{\mu_1}(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \alpha, \beta, \gamma_1)$$

where

$$h_g(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \beta, \alpha) := \frac{Z_i \epsilon_{S_i} h_g^C(X_i, S_i, \tilde{T}_i, \Delta_i; \beta)}{\pi(X_i, Z_i, A_i, \tilde{T}_i, \Delta_i; \alpha)}$$

$$h_{\mu_0}(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \alpha, \beta, \gamma_0) := \frac{(1 - Z_i)(1 - A_i)\epsilon_{S_i} h_{\mu_0}^C(g(X_i, S_i, \tilde{T}_i, \Delta_i; \beta), X_i; \gamma_0)}{\pi(X_i, Z_i, A_i, \tilde{T}_i, \Delta_i; \alpha)}$$

$$h_{\mu_1}(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \alpha, \beta, \gamma_1) := \frac{(1 - Z_i) A_i \epsilon_{S_i} h_{\mu_1}^C(g(X_i, S_i, \tilde{T}_i, \Delta_i; \beta), X_i; \gamma_1)}{\pi(X_i, Z_i, A_i, \tilde{T}_i, \Delta_i; \alpha)}.$$

The plug in estimators $\widehat{\theta}_{0,\text{plug-in}}$ and $\widehat{\theta}_{1,\text{plug-in}}$ can be written as the solutions ϕ_0, ϕ_1, respectively, of the following estimating equations:

$$0 = \sum_{i=1}^{n} (1 - Z_i)\{\mu_0(X_i; \gamma_0) - \phi_0\}$$

$$0 = \sum_{i=1}^{n} (1 - Z_i)\{\mu_1(X_i; \gamma_1) - \phi_1\}.$$

We define the parameter vector $\nu := (\beta, \gamma_0, \gamma_1, \alpha, \phi_0, \phi_1)$. Our final set of stacked estimating functions are

$$h_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \nu) = \begin{pmatrix} h_\pi(\epsilon_{Si}, X_i, Z_i, A_i, \tilde{T}_i, \Delta_i; \alpha) \\ h_g(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \beta, \alpha) \\ h_{\mu_0}(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \alpha, \beta, \gamma_0) \\ h_{\mu_1}(X_i, S_i, Z_i, A_i, \epsilon_{S_i}, \tilde{T}_i, \Delta_i; \alpha, \beta, \gamma_1) \\ (1 - Z_i)\{\mu_0(X_i; \gamma_0) - \phi_0\} \\ (1 - Z_i)\{\mu_1(X_i; \gamma_1) - \phi_1\} \end{pmatrix}$$

along with the stacked estimating equation

$$0 = \sum_{i=1}^{n} h_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \nu).$$

From this estimating equation, we obtain an estimator $\hat{\nu}$. Now, we define the following:

$$h'_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \nu) = \partial h_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \nu)/\partial \nu,$$

$$W_n\left(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{-h'_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu})\right\}$$

$$Q_n\left(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}\right) = \frac{1}{n}\sum_{i=1}^{n} h_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}) h_{\text{stack}}\,(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu})^T.$$

The sandwich variance estimator for $\hat{\nu}$ is

$$V_n\left(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}\right) = W_n\left(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}\right)^{-1} Q_n\left(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}\right) \left\{W_n\left(X_i, S_i, Z_i, A_i, \epsilon_{Si}, \tilde{T}_i, \Delta_i; \hat{\nu}\right)^{-1}\right\}^T.$$

E. Application of the Provisional Approval Causal Roadmap to the GBS Vaccine Development Case Study: Steps 1–6

E.1 *Step 1: Specify the causal model based on available knowledge of the context and proposed study*

The concentration of IgG antibodies that bind to various GBS surface proteins is highly predictive of IGbsD in many natural history studies. The OpKA functional assay is believed to measure a causal mechanism of vaccine protection. Thus, some subject matter experts suggest the perfect-surrogate causal model of Figure 2 [Panel (A)] approximately holds. However, one reason the imperfect-surrogate causal model [Panel (B)] is more appropriate is that IgG and OpKA levels for vaccinated and not infected with GBS vs. naturally GBS infected mothers may have different time-patterns in the infant from pre-birth through 90 days of age. A second reason is that the vaccine exposes a mother and baby to a specific GBS protein whereas natural-infection exposes a mother and baby to the entire GBS bacterium, meaning that natural-infection induces additional immune responses, and, if these responses help protect against IGbsD, then the bias function $u^{CT}(X, S)$ would likely be positive.

E.2 *Step 2: Define the causal parameter of interest*

The causal parameter of interest is $\text{VE} = 1 - E[Y(1)|Z = 0]/E[Y(0)|Z = 0]$ with $Y = I(T \leqslant 90)$: vaccine efficacy against IGbsD occurrence through $t_0 = 90$ days of life in the population of live-born infants with GBS colonized mothers.

E.3 *Step 5: Estimate the statistical estimand*

We consider the plug-in estimator and the one-step estimator of $E[Y(0)|Z = 0]$, $E[Y(1)|Z = 0]$, and the target parameter TE = VE. For estimating $g(X, S)$ from the sero-epidemiological observational studies, the baseline variables X are taken to be a set of known prognostic risk factors for IGbsD: X_1 the indicator of gestational age less than 37 weeks, and X_2 maternal age in years where younger age is a risk factor. The putative surrogate S is the average $\log_{10}$ IgG concentration against the four GBS alpha types in the vaccine. As noted in Section 3.4.3, many different IPS-weighted regression estimators for $g(X, S)$ may be considered. A superlearner estimator has appeal for providing an estimated optimal surrogate (EOS) to use in the estimation formulas, where several EOSs would be developed each under a different set of input variables (X, S), providing an empirical approach to selecting a best-predictive and parsimonious (X, S) to include in the phase 3 surrogate endpoint study data analysis (Price *and others*, 2018). For specification of the phase 3 statistical analysis plan, it would be useful to conduct a simulation study that compares performance of different implementations of the estimators (and variance estimators), not only considering different input variable sets but also different libraries of learners in the superlearner and different implementation details such as loss function and cross-validation scheme.

E.3.1 *Intercurrent events* The set of eligible participants for inclusion in the sero-epidemiological study data analysis are live-born infants of mothers colonized with GBS. As such, the methods provide estimation of VE for this colonized population. However, the public health goal for the indication of the vaccine is to include all pregnant persons irrespective of GBS colonization, avoiding the need for GBS colonization screening prior to vaccination. Therefore, an important goal is to infer that VE for the entire randomized cohort of pregnant mothers is sufficiently high. If the probability of IGbsD is zero for infants born to uncolonized mothers, then this inference is fairly

straightforward. Let $\text{TE}^{\text{Colon.}}$ be treatment efficacy against the target outcome in the colonized subgroup and let TE.against.Colon. be treatment efficacy against maternal GBS colonization by time τ, i.e., TE.against.Colon. $= 1 - P(C(1) = 1|Z = 0)/P(C(0) = 1|Z = 0)$ with $C(a)$ the potential outcome indicator of whether the mother is colonized by τ under randomization assignment a, for $a = 0, 1$. Then

$$\text{TE} = 1 - (1 - \text{TE}^{\text{Colon.}})(1 - \text{TE.against.Colon.}), \tag{E.1}$$

which implies that as long as the vaccine does not increase the rate of colonization, then vaccine efficacy in the randomized cohort will be at least as high as that in the colonized. Vaccine efficacy against GBS colonization can be readily estimated in the phase 3 study, such that the above formula can be used to seek assurance that focusing on estimating TE in the colonized supports sufficient treatment efficacy for the whole randomized cohort.

Indeed, the risk of IGbsD in uncolonized mothers is close to zero (Vekemans *and others*, 2019) (some would argue equal to zero), which is the reason the sero-epidemiological studies restrict the study cohort to colonized mothers. In fact, being uncolonized means that the bacteria are not alive and growing in which case IGbsD should be impossible; however, colonization is measured by a diagnostic procedure applied at one or a discrete set of sampling time points, and imperfect sensitivity of the colonization diagnostic test could imply low residual risk of IGbsD in infants born from mothers diagnosed as uncolonized. Missing detection of colonized GBS could also occur due to post partum GBS acquisition and non-vertical GBS transmission to the infant.

If there is concern of non-negligible risk of IGbsD for infants born to mothers diagnosed as uncolonized, then an approach to inferring sufficient VE for the whole randomized cohort is to rely on the estimate of VE for the colonized sub-population, supplemented by arguments that this estimate of VE serves as a lower bound of the estimate of VE for the whole randomized cohort. One source of evidence would be data showing that the GBS vaccine reduces the rate of colonization in the randomized cohort, or at least does not increase the rate, as noted above. This can readily

be studied as an objective in the phase 3 trial. A second source of evidence would be data supporting that VE against IGbsD in uncolonized mothers is at least as high as VE against IGbsD in colonized mothers. In principle, argumentation for this point could be based on the sampling design for measurement of S in the phase 3 study, measuring S for both colonized and uncolonized participants, from both the vaccine and placebo arms, and applying the VE estimation formula to both subgroups. However, because the estimator $\hat{g}(X, S)$ would be different for the colonized vs. uncolonized groups, this approach would only be viable if the observational study includes the uncolonized group in the correlates of risk model building. This is not possible for the real GBS studies given zero probability of sampling S for the uncolonized group. Some less direct evidence could be generated from the phase 3 trial in the form of comparing the distribution of S between the colonized vs. uncolonized groups within each of the vaccine and placebo arms. The assumption of vaccine efficacy being as least as high in uncolonized mothers may be reasonable based on the observation that vaccines typically provide better efficacy when the amount of pathogen exposure (e.g., number of virus copies is less) (Kaslow, 2021). Counterbalancing this argument, an immunological reason why the colonized could have have higher efficacy is that colonization generates immune responses and the vaccination serves as a boost, whereas the uncolonized may be naive to the pathogen such that the vaccination is like a prime, and typically vaccines confer greater efficacy in groups with prior infection compared to in infection-naive groups [e.g., Capeding *and others* (2014); Villar *and others* (2015); Sadoff *and others* (2022); Dayan *and others* (2023*b*,*a*)].

In the GBS application, $Y^{0-\tau}$ is the indicator of acquiring IGbsD by birth, such that the simplifying assumption $P(Y^{0-\tau} = 1) = 0$ amounts to no in-utero IGbsD that has onset on the date of birth. If such events do occur, then additional investigations would be warranted to provide evidence for whether restricting the analysis to infants with $Y^{0-\tau} = 0$ would be expected to affect VE in the whole cohort. An alternative framing would redefine $Y^{0-\tau} = 1$ to mean

IGbsD onset *before* birth and therefore structurally to make the assumption $P(Y^{0-\tau} = 1) = 0$ hold, in which case S measured from cord blood would need to be considered definable, where some infants may have IGbsD onset at birth that are assigned a failure time of one day. This latter framing may be reasonable based on the fact that infants receive the S immune marker levels passively from the mother.

E.4 *Step 6: Quantify the uncertainty in the estimate of the statistical estimand*

It is of interest to compare the two variance estimation approaches for the plug-in estimator (sandwich, bootstrap) for the specific context of the planned/available GBS sero-epidemiological studies and potential phase 3 study designs. These two variance estimators were evaluated in Section 3.5. It is also of interest to study the strategies for specifying the bias functions considered in Section 3.5.1.

F. Simulation study details and additional simulation studies for the GBS vaccine development case study

F.1 *Design of simulation study conditions*

The simulation study was designed to approximately match published characteristics of GBS and its risk factors.

1. Probability of IGbsD by 90 days of age for colonized mothers ≈ 0.005. From Vekemans *and others* (2019), approximately 20% of mothers are colonized and 0.1% of all infants have IGbsD. We obtain an approximate incidence of 0.005 by assuming that only infants of colonized mothers can develop invasive GBS disease.

2. Geometric mean (95% CI) of cord-blood IgG concentration 0.01 (0.01-0.02) in IGbsD cases and 0.04 (0.03 - 0.06) in controls, which was observed for infant RibN IgG in Dangor *and*

others (2023).

3. Covariates X_1, X_2, X_3, where X_1 represents pre-term birth (less than 37 weeks gestational age), X_2 represents maternal age (younger age a risk factor), and X_3 represents a continuous covariate unrelated to the outcome.

$$X_1|Z, A \sim \text{Bernoulli}(0.05), X_2|Z, A \sim \text{Uniform}(18, 40), X_3|Z, A \sim \text{Normal}(0, 1)$$

   - Pre-term birth is associated with roughly 2.56× odds of early onset IGbsD (Puopolo *and others*, 2011).
   - Maternal age < 25 years is associated with roughly 1.94× odds of IGbsD (Parente *and others*, 2017).

Based on these constraints, we set the following to be our true $A = 0$ data generating conditional regression function:

$$P(Y = 1|S, X_1, X_2, X_3, A = 0, Z) = \beta_0 + \beta_1 S + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3$$

where $\beta_0 = -17.1$, $\beta_1 = -8.2$, $\beta_2 = 0.69$, $\beta_3 = -0.03$, $\beta_4 = 0$, and we set our distribution of $S \sim \text{Normal}(-1.45, 0.0225)$, where S represents the log IgG biomarker, $X_1 \sim \text{Bernoulli}(0.05)$, $X_2 \sim \text{Uniform}(18, 40)$, and $X_3 \sim \text{Normal}(0, 1)$.

With this data-generating function, we have that in our observational study (using empirical measures from a simulated n =10,000,000 size data set):

- Incidence of IGbsD by 90 days ≈ 0.005
- Geometric mean cord-blood IgG biomarker in IGbsD cases is 0.01 and in controls is 0.04
- Pre-term birth associated with 2× odds of disease
- Each one-year increase in maternal age associated with 3% lower odds of disease

Table 1. Results for the Supplemental Simulation Study with approximately 3 times higher IGbsD outcome rate (expect 585 invasive GBS disease cases by 90 days of age) on empirical bias, median standard error, standard deviation of estimates, and 95% confidence interval coverage of $E[Y(0)|Z=0]$, $E[Y(1)|Z=0]$, and VE. SE (bs) = bootstrap standard error, SE (sw) = sandwich standard error, SD = standard deviation of estimates, Cov = 95% confidence interval coverage using sandwich standard error. For the VE entries the standard errors shown are for $\log(1-\widehat{\text{VE}})$. The S data are sampled from 250 participants in each treatment arm of the phase 3 study.

	Plug-In Estimator					One-Step Estimator			
	Bias	SE (bs)	SE (sw)	SD	Cov	Bias	SE (os)	SD	Cov
True VE = 0									
$E[Y(0)\|Z=0]$	0	0.00136	0.00146	0.00147	0.95	0	0.00136	0.00146	0.94
$E[Y(1)\|Z=0]$	-0.000050	0.00136	0.00146	0.00147	0.96	-0.00010	0.00136	0.00145	0.94
VE	0.00623	0.117	0.115	0.121	0.95	0.00346	0.11	0.12	0.94
True VE = 0.5									
$E[Y(0)\|Z=0]$	-0.00010	0.00136	0.00146	0.0015	0.96	-0.00010	0.00136	0.0015	0.94
$E[Y(1)\|Z=0]$	-0.000060	0.00101	0.00103	0.0011	0.93	-0.00015	0.00101	0.00115	0.92
VE	0	0.152	0.148	0.154	0.94	0.006	0.15	0.161	0.95
True VE = 0.9									
$E[Y(0)\|Z=0]$	-0.00010	0.00136	0.00146	0.00148	0.94	-0.00020	0.00135	0.00149	0.93
$E[Y(1)\|Z=0]$	-0.000040	0.000255	0.00026	0.000292	0.92	0.00011	0.000361	0.000487	0.93
VE	0.003	0.193	0.189	0.202	0.95	-0.0085	0.233	0.283	0.89

F.2 *Results from the larger sample simulation study*

In this section, we report results for a simulation study with a higher case rate. We set the following to be our true $A=0$ data generating conditional regression function:

$$P(Y=1|X_1,X_2,X_3,A=0,Z,S)=\beta_0+\beta_1 S+\beta_2 X_1+\beta_3 X_2+\beta_4 X_3$$

where $\beta_0=-14$, $\beta_1=-7$, $\beta_2=0.69$, $\beta_3=-0.03$, $\beta_4=0$.

We preserve the same distribution of S as in Simulation Study 1, with $S\sim\text{Normal}(-1.45,0.225)$, resulting in a baseline case rate of 0.016 (about 3 times higher than 0.005).

To generate VE $=\{0,0.5,0.9\}$, we manipulate the distribution of the biomarker in the vaccine arm $S|A=1,Z=0$:

- To create VE = 0, set $S|A=1,Z=0\sim\text{Normal}(-1.45,0.0225)$
- To create VE = 0.5, set $S|A=1,Z=0\sim\text{Normal}(-1.29,0.04)$
- To create VE = 0.9, set $S|A=1,Z=0\sim\text{Normal}(-1.04,0.0441)$.

Finally, matching the simulations in the main article, we sample 250 participants from each

of the vaccine and placebo arms for measurement of S in the phase 3 study ($Z = 0$).

Results are shown in Table 1. In this higher case rate scenario, we observe minimal bias in estimating VE, approximately nominal confidence interval coverage, and close agreement between bootstrap, sandwich, and empirical standard errors.

References

Athey, S, Chetty, R, Imbens, GW and Kang, H. (2024). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *National Bureau of Economic Research Working Papers* **Working Paper 26463 updated April 2024**.

Capeding, Maria Rosario, Tran, Ngoc Huu, Hadinegoro, Sri Rezeki S, Ismail, Muhammad, Chotpitayasunondh, Tawee, Chua, Mary Noreen, Luong, Chan Quang, Rusmil, Kusnandi, Wirawan, Dewa Nyoman, Nallusamy, Revathy, Pitisuttithum, Punnee, Thisyakorn, Usa, Yoon, In-Kyu, van der Vliet, Diane, Langevin, Edith, Laot, Thelma, Hutagalung, Yanee, Frago, Carina, Boaz, Mark, Wartel, T Anh, Tornieporth, Nadia G, Saville, Melanie, Bouckenooghe, Alain *and others*. (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**(9951), 1358–1365.

Dangor, Ziyaad, Kwatra, Gaurav, Pawlowski, Andrzej, Fisher, Per B, Izu, Alane, Lala, Sanjay G, Johansson-Lindbom, Bengt and Madhi, Shabir A. (2023). Association of infant Rib and Alp1 surface protein N-terminal domain immunoglobulin G and invasive Group B Streptococcal disease in young infants. *Vaccine* **41**(10), 1679–1683.

Dayan, Gustavo H, Rouphael, Nadine, Walsh, Stephen R, Chen, Aiying, Grunenberg, Nicole, Allen, Mary, Antony, Johannes, Asante, Kwaku Poku, Bhate,

Amit Suresh, Beresnev, Tatiana *and others*. (2023*a*). Efficacy of a bivalent (d614+ b. 1.351) sars-cov-2 recombinant protein vaccine with as03 adjuvant in adults: a phase 3, parallel, randomised, modified double-blind, placebo-controlled trial. *The Lancet Respiratory Medicine* **11**(11), 975–990.

Dayan, Gustavo H, Rouphael, Nadine, Walsh, Stephen R, Chen, Aiying, Grunenberg, Nicole, Allen, Mary, Antony, Johannes, Bhate, Amit Suresh, Beresnev, Tatiana, Bonaparte, Matthew I *and others*. (2023*b*). Efficacy of a monovalent (d614) sars-cov-2 recombinant protein vaccine with as03 adjuvant in adults: a phase 3, multi-country study. *Eclinicalmedicine* **64**.

Kaslow, David C. (2021). Force of infection: a determinant of vaccine efficacy? *npj Vaccines* **6**(1), 51.

Parente, Victoria, Clark, Reese H, Ku, Lawrence, Fennell, Courtney, Johnson, Makaela, Morris, Emma, Romaine, Andrew, Utin, Uty, Benjamin, Daniel K, Messina, Julia A *and others*. (2017). Risk factors for group b streptococcal disease in neonates of mothers with negative antenatal testing. *Journal of Perinatology* **37**(2), 157–161.

Price, Brenda L, Gilbert, Peter B and van der Laan, Mark J. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics* **74**(4), 1271–1281.

Puopolo, Karen M, Draper, David, Wi, Soora, Newman, Thomas B, Zupancic, John, Lieberman, Ellice, Smith, Myesha and Escobar, Gabriel J. (2011). Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. *Pediatrics* **128**(5), e1155–e1163.

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

Sadoff, Jerald, Gray, Glenda, Vandebosch, An, Cárdenas, Vicky, Shukarev, Georgi, Grinsztejn, Beatriz, Goepfert, Paul A, Truyers, Carla, Van Dromme, Ilse, Spiessens, Bart *and others*. (2022). Final analysis of efficacy and safety of single-dose Ad26.COV2.S. *New England Journal of Medicine* **386**(9), 847–860.

Shepherd, B, Gilbert, Peter B, Jemiai, Y and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62**, 332–342.

Stefanski, Leonard A and Boos, Dennis D. (2002). The calculus of M-estimation. *The American Statistician* **56**(1), 29–38.

Vekemans, Johan, Moorthy, Vasee, Friede, Martin, Alderson, Mark R, Sobanjo-Ter Meulen, Ajoke, Baker, Carol J, Heath, Paul T, Madhi, Shabir A, Mehring-Le Doare, Kirsty, Saha, Samir K *and others*. (2019). Maternal immunization against Group B Streptococcus: World Health Organization research and development technological roadmap and preferred product characteristics. *Vaccine* **37**(50), 7391–7393.

Villar, Luis, Dayan, Gustavo Horacio, Arredondo-García, José Luis, Rivera, Doris Maribel, Cunha, Rivaldo, Deseda, Carmen, Reynales, Humberto, Costa, Maria Selma, Morales-Ramírez, Javier Osvaldo, Carrasquilla, Gabriel, Rey, Luis Carlos, Dietze, Reynaldo, Luz, Kleber, Rivas, Enrique, Montoya, Maria Consuelo Miranda, Supelano, Margarita Cortés, Zambrano, Betzana, Langevin, Edith, Boaz, Mark, Tornieporth, Nadia, Saville, Melanie *and others*. (2015). Efficacy of a tetravalent dengue vaccine in children in Latin America. *New England Journal of Medicine* **372**, 113–123.