Journal de la Société Française de Statistique
Vol. 161 No. 1 (2020)

Assessing trends in vaccine efficacy by pathogen genetic distance

David Benkeser[1] , Michal Juraska[2] and Peter B. Gilbert[2]

Abstract: Preventive vaccines are an effective public health intervention for reducing the burden of infectious diseases, but have yet to be developed for several major infectious diseases. Vaccine sieve analysis studies whether and how the efficacy of a vaccine varies with the genetics of the infectious pathogen, which may help guide future vaccine development and deployment. A standard statistical approach to sieve analysis compares the effect of the vaccine to prevent infection and disease caused by pathogen types defined dichotomously as genetically near or far from a reference pathogen strain inside the vaccine construct. For example, near may be defined by amino acid identity at all amino acid positions considered in a multiple alignment and far defined by at least one amino acid difference. An alternative approach is to study the efficacy of the vaccine as a function of genetic distance from a pathogen to a reference vaccine strain where the distance cumulates over the set of amino acid positions. We propose a nonparametric method for estimating and testing the trend in the effect of a vaccine across genetic distance. We illustrate the operating characteristics of the estimator via simulation and apply the method to a recent preventive malaria vaccine efficacy trial.

Keywords: vaccines, competing risks, causal inference, marginal structural model, Hamming distance
AMS 2000 subject classifications: 62G10, 62N03, 62P10

1. Introduction

Over the past century, disease burden due to infectious pathogens has been substantially reduced by preventive vaccines. However, many existing vaccines are only partially efficacious, a fact that may be explained in part by genetic heterogeneity of pathogens. Whereas vaccines are typically constructed using only one or several specific pathogen sequences, pathogens may exhibit broad genetic heterogeneity. Thus, the vaccine may stimulate immune responses that are protective against infection or disease caused by pathogens with these few sequences, but may not confer protection more broadly, leading to reduced overall efficacy. Therefore, it is often informative to study whether and how the efficacy of a partially effective vaccine varies by pathogen genetics.

A common analogy used to describe the protective mechanism of vaccines is to imagine a vaccine as a sieve: the vaccine blocks infection or disease caused by some genotypes of pathogens, but lets others through. Equivalently, the sieve may be regarded as the latent immune response elicited by vaccination, which may differ based on pathogen genotypes and may impact risk of infection or disease. These analogies have led to the study of vaccine efficacy as a function of pathogen genetics to be referred to as *vaccine sieve analysis* (Gilbert et al., 1998, 2001). A vaccine is said to exhibit a *sieve effect* at a particular genetic region of the pathogen if the vaccine is differentially efficacious against pathogens depending on their amino acid sequence in that region. Identification of sieve effects may help guide the selection of antigens to include in future multivalent vaccines.

It is common practice in sieve analysis to compare binary categories of genetic sequences, such as the vaccine's efficacy for preventing infection or disease caused by (a) pathogens that are *fully matched* to the vaccine in a particular region, i.e., have the same genetic sequence versus (b) pathogens that are *mismatched* to the vaccine in this region, i.e., have at least one amino acid difference, (e.g., Rolland et al. (2012); Neafsey et al. (2015)). We expect that if a sieve effect is present, (a) will be greater than

[1] Department of Biostatistics and Bioinformatics, Emory University; 1518 Clifton Rd. NE; Atlanta, GA USA 30322
E-mail: benkeser@emory.edu
[2] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center; 1100 Fairview Ave. N; Seattle, WA USA 98109
E-mail: mjuraska@fhcrc.org and E-mail: pgilbert@scharp.org

 ISSN: 2102-6238

(b), indicating that the vaccine works better against matched than mismatched pathogens. Under a binary categorization, many genotypes are combined in the mismatched category: the category includes pathogens with a single amino acid that differs from the vaccine, as well as pathogens with many amino acids that differ from the vaccine. An alternative approach is to study how vaccine efficacy depends on *genetic distance* from the vaccine, which can be helpful for better understanding mechanisms of protection associated with antibody epitopes in the selected region. This approach can help overcome the common problem that, although a large number of discrete genotypes exist, there are often too-few infection or disease endpoints of individual genotypes to reliably assess vaccine efficacy against each genotype. Several methodologies have been proposed for estimating genotype-specific vaccine efficacy with genotypes defined by genetic distance from the vaccine (Gilbert et al., 2008; Sun et al., 2009, 2012; Juraska and Gilbert, 2013, 2016). Whereas most approaches have developed tests centered around parameters that describe the vaccine's effect on the hazard or *instantaneous risk* of disease or infection acquisition, Gilbert et al. (2008) studies the vaccine's effect on *cumulative risk* of disease or infection as a function of genetic distance, which may be more relevant for public health decision making in the common setting of waning vaccine efficacy. We propose a new approach to studying the vaccine's effect on *cumulative risk* as a function of genetic distance. Our approach differs from that of Gilbert et al. (2008) in several ways. Gilbert and colleagues focused on nonparametric smoothing for a continuous distance, while our approach treats distance as an ordered count with many categories and avoids nonparametric smoothing. Our approach allows for participant dropout to depend on measured characteristics, whereas the approach of Gilbert and colleagues assumed random censoring. Finally, our approach lends itself to a causal interpretation under assumptions, whereas that of Gilbert et al. (2008) does not.

Our study is motivated by a Phase III randomized, controlled trial of the RTS,S/AS01 vaccine (hence, RTS,S), which, from a regulatory perspective, is the most advanced vaccine candidate for malaria prevention. Malaria is the clinical disease associated with infection by the *Plasmodium* parasite. The parasite has a complicated life-cycle, which involves stages in the human blood and liver, as well as stages in a mosquito host. The RTS,S vaccine consists of a single antigen: a portion of the circumsporozoite protein (CS protein) found on the surface of the *Plasmodium falciparum* parasite during the human blood phase of its life-cycle. The vaccine reduced the average instantaneous risk of clinical malaria over 12 months by an estimated 63% in 5–17 month old children (RTS,S Clinical Trials Partnership, 2011; Agnandji et al., 2012) and pilot implementation programs are disseminating the vaccine in areas of high malaria transmission (WHO, 2016). The protective mechanism of the vaccine is incompletely understood and differing hypotheses exist regarding how immunity is mediated through immune responses. We are thus motivated to study how vaccine efficacy varies as a function of genetic distance to contribute to scientific understanding of the immune mechanisms associated with the RTS,S vaccine.

The remainder of the article is organized as follows. In Section 2, we introduce notation, formalize our definition of a sieve effect, and define a parameter to describe the trend in vaccine efficacy as a function of genetic distance. In Section 3, we discuss estimation of sieve effects and inference for the trend parameter. In Section 4, we include a simulation study and in Section 5 we analyze the RTS,S data. We conclude with a discussion.

2. Notation and parameter of interest

Data typical of vaccine sieve analysis are generated as follows. Trial participants are enrolled and baseline covariates W are measured. Participants are subsequently randomly assigned to either an active vaccine ($Z = 1$) or a control vaccine ($Z = 0$). The random assignment could depend on covariates, as in the RTS,S/AS01 trial, where vaccine assignment was randomized within each of eleven study sites. Participants are followed for a fixed study period and monitored for the occurrence of a study endpoint, such as pathogen infection or clinical disease with documentation of pathogen infection. We use T to denote the time from baseline until the first endpoint. We assume that T takes only a finite-number of values, as is the case if T corresponds to days until first endpoint. Participants who do not experience an endpoint are followed until completion of the study at a pre-specified time τ. It is common in trials with

 ISSN: 2102-6238

longitudinal follow-up that some participants are right-censored before τ. We use U to denote the time until last follow-up. We set $U = \tau$ for participants who do not experience a study endpoint by τ. For each observed endpoint, we obtain the genetic sequence of the pathogen at the time of the endpoint. Based on a multiple alignment of pathogen amino acid sequences from subjects with the failure event, we use $J \in \{1, \ldots, K+1\}$ to categorize sequences based on their genetic distance from the vaccine antigen in a genetic region of interest comprised of K amino acids. Sequences are categorized so that a categorization of $J-1$ corresponds to endpoints with genetic distance J. For example, we use $J = 1$ to denote sequences that are fully matched to the vaccine along the entire region of interest (i.e., genetic distance of zero), $J = 2$ for sequences with genetic distance of one, and so on. Due to censoring, we do not observe T, U, and J for all participants; instead, we observe $\tilde{T} := \min(T, U)$ and ΔJ, where $\Delta := I(\tilde{T} = T)$. The observed data are assumed to be independent and identically distributed copies of $O := (W, Z, \tilde{T}, \Delta, \Delta J)$.

For our developments, it is useful to alternatively express the observed data in terms of discrete counting processes. Specifically, we write $O = \big(W, Z, \{N_j(t), C(t) : j = 0, \ldots, K \text{ and } t = 1, ..., \tau\}\big)$, where for $j = 0, \ldots, K$, $N_j(t) := I(\tilde{T} \leq t, \Delta J = j+1)$, and $C(t) := I(\tilde{T} \leq t, \Delta = 0)$. In words, $N_j(t)$ is an indicator that an observed endpoint occurred prior to or at time t, and that the pathogen associated with endpoint had $j+1$ distance from the vaccine insert. Similarly, $C(t)$ is an indicator that right-censoring has occurred prior to or at time t. We will also use the shorthand $N_\cdot(t)$ to simultaneously refer to $N_j(t)$ for all $j = 0, \ldots, K$; for example, we write $N_\cdot(t) = 0$ to denote that $N_0(t) = 0, \ldots, N_K(t) = 0$. By convention, if a participant has an endpoint with genetic distance k at time s, then we set $N_k(t) = 1$ for all $t > s$, $N_j(t) = 0$ for $j \neq k$ and $t > s$, and $C(t) = 0$ for all $t \geq s$. If a participant is censored at time s, we arbitrarily set $N_\cdot(t) = 0$ for all $t > s$. We denote by P_0 the true distribution of O and denote by $\mathscr{M}$ our statistical model, which we take to be nonparametric. However, our theoretical developments apply to any model that makes assumptions about the probability of receiving vaccine given covariates and the probability of censoring given vaccine status and covariates.

We define our causal parameter of interest using a structural causal model (Pearl, 2009). We assume that each component of the observed data structure is a function of a set of observed parent variables and an unmeasured exogenous error term. The observed parent of Z is W. For $t = 1, \ldots, \tau$, the observed parents of $N_j(t)$ are W, Z, $C(t-1)$, and $N_\cdot(t-1)$, and the observed parents of $C(t)$ are W, Z, $C(t-1)$ and $N_\cdot(t)$. We denote by $\mathbb{P}_0^z$ the distribution the data would have under an intervention that sets $Z = z$ (e.g., assigning vaccine to all participants) and sets $C(t) = 0$ for $t = 1, \ldots, \tau - 1$ (i.e., assigning all participants to remain under study). We refer to this distribution as the post-intervention distribution and define $N_\cdot^z(t)$ to be a counterfactual random variable with this distribution. The counterfactual parameter

$$F_{j,0}^z(\tau) := E_{\mathbb{P}_0^z}\{N_j^z(\tau)\} \tag{1}$$

is the proportion of participants who experience an endpoint with genetic distance j from the vaccine antigen by time τ if all participants are assigned $Z = z$ and remain under study until τ. A typical summary of the causal effect of the vaccine relative to the control vaccine for preventing endpoints with genetic distance j is genotype-specific vaccine efficacy $\text{VE}_{j,0}(\tau) := 1 - F_{j,0}^1(\tau)/F_{j,0}^0(\tau)$. Values of $\text{VE}_{j,0}(\tau)$ near to one indicate reduced incidence of endpoints caused by a pathogen of genetic distance j. Note that the scale of $\text{VE}_{j,0}(\tau) \in (-\infty, 1]$ is not symmetric, so we instead focus on a parameter with a symmetric scale, the log-ratio of counterfactual cumulative incidence control vs. vaccine,

$$L_{j,0}(\tau) := \log\left\{\frac{F_{j,0}^0(\tau)}{F_{j,0}^1(\tau)}\right\}. \tag{2}$$

Values of $L_{j,0}(\tau)$ greater than zero indicate a higher incidence of j-distance endpoints when control vaccine is assigned compared to when the active vaccine is assigned. Large positive values of $L_{j,0}(\tau)$ indicate the vaccine works well at preventing endpoints with distance j from the vaccine. Furthermore, $L_{j,0}(\tau)$ can be related back to vaccine efficacy, $\text{VE}_{j,0}(\tau) = 1 - \exp\{-L_{j,0}(\tau)\}$.

We are interested in testing the null hypothesis that $L_{j,0}(\tau)$ is the same for all $j = 0, \ldots, K$. One approach for testing this hypothesis is to estimate $L_{j,0}(\tau)$ for $j = 0, \ldots, K$, test each pairwise null hypothesis $L_{j,0} = L_{k,0}$ for $j = 0, \ldots, K$ and $k \neq j$, and perform a multiplicity correction to properly control

 ISSN: 2102-6238

the type-one error rate. However, if K is large, there will be many pairwise comparisons and multiplicity corrections may result in a test with low overall power to detect sieve effects. Instead, we propose to test for a trend in the effect across genetic distances. We expect that if there is a sieve effect present in the genetic region under study, then $L_{j,0}(\tau)$ will be monotone non-increasing in j. That is, the vaccine will work best against endpoints that are genetically similar to the vaccine antigen with efficacy decreasing as endpoints become less similar to the vaccine. Thus, we would like to design a test that has high power under this alternative hypothesis. To that end, we consider the parameter

$$(\alpha_0, \beta_0) := \text{argmin}_{(\alpha,\beta)} \sum_{j=0}^{K} \omega_j \left\{ L_{j,0}(\tau) - \alpha - \beta j \right\}^2, \tag{3}$$

where $\omega := (\omega_j : j)$ is a user-specified vector of positive weights. We discuss choices of weight function in Section 3.4. This parameter equals the weighted L^2 projection of the true function describing how $L_{j,0}(\tau)$ varies with j onto a linear working model. Notice that the linear working model is truly a working model in the sense that the definition of the parameter does not depend on the true function being linear. The parameter β_0 merely provides a useful summary of the trend in the vaccine's effect across genetic distance. In particular, note that $\beta_0 = 0$ corresponds with the null hypothesis of interest, that $L_{j,0}(\tau)$ is the same for all $j = 0, \ldots, K$, while values of β_0 less than 0 indicate decreasing efficacy with increasing distance. Our goal is to estimate β_0 and design a test of the null hypothesis that $\beta_0 = 0$.

3. Identification, estimation and inference

3.1. Identification

Our approach to estimating β_0 is to estimate the counterfactual cumulative incidence $F^z_{j,0}(\tau)$ for $z = 0, 1$ and $j = 0, \ldots, K$. Subsequently, these estimates are plugged into (2) and (3) to obtain estimates of $L_{j,0}$ and (α_0, β_0), respectively. The counterfactual cumulative incidence $F^z_{j,0}(\tau)$ may be estimated using the observed data under the following assumptions:

1. (consistency) $N^z_j(t) = N_j(t)$ if $Z = z, C(t-1) = 0$ for $t = 1, \ldots, \tau - 1$;
2. (no interference) the counterfactual outcome for participant i, $N^z_{j,i}(\tau)$ depends only on the treatment assignment for patient i;
3. (sequential randomization) $N^z_j(t) \perp Z \mid W$ and $N^z_j(t) \perp C(t-1) \mid Z = z, N_\cdot(t-1), W$, for $t = 1, \ldots, \tau$;
4. (positivity) $P_0(Z = z \mid W)$ and $P_0(C(t-1) = 0 \mid Z = z, C(t-2) = 0, N_\cdot(t-2) = 0, W)$ for $t = 1, \ldots, \tau - 1$ are each strictly greater than zero with probability one.

The first two assumptions are fundamental in order to ensure that the counterfactual endpoints are well defined. The consistency assumption essentially says that the hypothetical intervention that assigns vaccine and no censoring does not fundamentally alter the way the vaccine works. Thus, the outcome we see in the observed data is the same outcome we would have seen under this hypothetical intervention. The assumption of no interference is often dubious in infectious disease settings, where the infection status of a given participant might depend on whether or not their family and friends also received the vaccine (Hudgens and Halloran, 2008). Given the complicated life cycle of malaria, it is uncertain the degree to which the assumption of no interference may have been violated in the RTS,S trial. We proceed as though this assumption approximately holds, leaving to future work the development of methodologies which fully relax this assumption. The sequential randomization assumption states that there are no unmeasured confounders of vaccine assignment nor of censoring. The former is guaranteed by randomization of vaccine assignment in clinical trials, while the latter is generally not testable based on observed data. Instead, we must hope that the covariates W collected are sufficiently rich so as to include all possible variables that might influence both a participant's risk of the endpoint and her/his propensity to drop out of the trial. The positivity assumption states that there are no groups of participants

 ISSN: 2102-6238

with zero probability of receiving the vaccine and remaining uncensored. Because this is an assumption on the observed data distribution P_0, this assumption can be studied empirically (Petersen et al., 2010).

If these assumptions hold, then we can estimate counterfactual cumulative incidence by estimating

$$F^z_{j,0}(\tau) = \int \left(\sum_{t=1}^{\tau} \left[\lambda^z_{j,0}(t)(w) \prod_{s=1}^{t-1} \left\{ 1 - \lambda^z_{\cdot,0}(s)(w) \right\} \right] \right) dG_0(w) \,, \tag{4}$$

where for $t = 1, \ldots, \tau$, we define $\lambda^z_{j,0}(t)(w) := P_0(N_j(t) = 1 \mid Z = z, C(t-1) = 0, N_\cdot(t-1) = 0, W = w)$, $\lambda^z_{\cdot,0}(t)(w) := \sum_{j=0}^{K} \lambda_{j,0}(t)(w)$, and $G_0(w) := P_0(W \leq w)$. We refer to $\lambda^z_{j,0}$ as the conditional genotype-specific hazard function, $\lambda^z_{\cdot,0}$ as the conditional total hazard function, and G_0 as the distribution of baseline covariates. The key to proving (4) is that for $t = 1, \ldots, \tau$, under sequential randomization

$$\mathbb{P}^z_0(N^z_j(t) \mid Z = z, C(t-1) = 0, N^z_\cdot(t-1) = 0, W) = P_0(N(t) \mid Z = z, C(t-1) = 0, N_\cdot(t-1) = 0, W) \,. \tag{5}$$

That is, the covariate-conditional cause-specific hazard function for the counterfactual counting process is equal to the covariate-conditional cause-specific hazard function for the observed counting process. Intuitively, if W contains all information about meaningful differences between participants with respect to their probability of endpoints and censoring, then within each stratum defined by W at a given time, there are no meaningful differences (with respect to the probability of experiencing an endpoint) between participants who have previously dropped out of the trial and those who remain. Thus, the counterfactual probability of an endpoint at the next time point in each stratum is identical to the observed data probability of an endpoint at the next time point. Notice that the positivity assumption is required in order that the right-hand-side of (5) is well-defined.

Once the equivalence of counterfactual and observed data hazard functions is established, all that remains is to relate the observed data hazard function to cumulative incidence. Toward that end, it is helpful to enumerate the possible ways that an endpoint with genetic distance j may be observed by time τ given vaccine assignment $Z = z$ and covariates $W = w$. Such an endpoint may be observed at the first time point, which occurs with probability $\lambda^z_{j,0}(1)(w)$. The endpoint may also be observed to occur at the second time point, in which case no endpoint of any type must have occurred at the first time point. Given no endpoint at the first time point, the probability of an event at the second time point is $\lambda^z_{j,0}(2)(w)$, while the probability of no endpoint at the first time point is $1 - \lambda^z_{\cdot,0}(1)(w)$. Thus, the probability of observing a j-distance endpoint at the second time is $\lambda^z_{j,0}(2)(w)\{1 - \lambda^z_{\cdot,0}(1)(w)\}$ and the cumulative probability of observing an endpoint by the second time is $\lambda^z_{j,0}(1)(w) + \lambda^z_{j,0}(2)(w)\{1 - \lambda^z_{\cdot,0}(1)(w)\}$. The summation in (4) is thus made plain: the t-th term in the sum is the probability of observing a j-distance endpoint at t given no previous endpoint, $\lambda^z_{j,0}(t)(w)$, multiplied by the probability that no endpoint of any distance was observed prior to t, $\prod_{s=1}^{t-1}\{1 - \lambda^z_{\cdot,0}(s)(w)\}$. The sum therefore yields the cumulative probability of j-distance events in participants with $Z = z$ and $W = w$, while the integral averages these covariate-conditional probabilities with respect to the distribution of covariates in the participant population.

3.2. Estimation

Equation (4) implies that an estimate of cumulative incidence may be obtained by estimating $\lambda^z_{j,0}$, $\lambda^z_{\cdot,0}$, and G_0 and substituting these estimates into (4). If covariates are low-dimensional and discrete, then we can use empirical estimates for each of these components. For $t = 1, \ldots, \tau$ we define $dN^z_j(t)(w) := I(Z = z, C(t-1) = 0, N_\cdot(t-1) = 0, N_j(t) = 1, W = w)$ and $n^z_w(t) := \sum_{i=1}^{n} I(Z_i = z, C_i(t-1) = 0, N_{\cdot,i}(t-1) = 0, W_i = w)$. Empirical estimates of the conditional genotype-specific and total hazard at time $t = 1, \ldots, \tau$ can be computed respectively as

$$\lambda^z_{j,n}(t)(w) = \frac{1}{n^z_w(t)} \sum_{i=1}^{n} dN^z_{j,i}(t)(w) \quad \text{and} \quad \lambda^z_{\cdot,n}(t)(w) = \sum_{j=0}^{K} \lambda^z_{j,n}(t)(w) \,.$$

 ISSN: 2102-6238

Similarly, we may use empirical estimates of the distribution of baseline covariates,

$$G_n(w) := \frac{1}{n}\sum_{i=1}^{n} I(W_i = w) \,.$$

Together, these empirical estimates of the hazard functions and baseline covariate distribution give an estimate of cumulative incidence:

$$F^z_{j,n}(\tau) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{\tau}\left[\lambda^z_{j,n}(t)(W_i)\prod_{s=1}^{t-1}\left\{1-\lambda^z_{\cdot,n}(s)(W_i)\right\}\right]\right). \tag{6}$$

An estimate $L_{j,n}(\tau)$ of $L_{j,0}(\tau)$ may be computed for each $j = 0,\ldots,K$ by computing (6) for $j = 0,\ldots,K$ and $z = 0,1$ and substituting into (2). An estimate of (α_0,β_0) can be computed by substituting $L_{j,n}(\tau)$ into (3),

$$(\alpha_n,\beta_n) := \text{argmin}_{\alpha,\beta}\sum_{j=0}^{K}\omega_j\left\{L_{j,n}(\tau)-\alpha-\beta j\right\}^2 ,$$

which can be computed explicitly. To wit, we define

$$\boldsymbol{X} := \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & K \end{pmatrix}, \quad \boldsymbol{\Omega} := \begin{pmatrix} \omega_0 & 0 & \ldots & 0 \\ 0 & \omega_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \omega_K \end{pmatrix}, \text{ and } \boldsymbol{L}_n := \begin{pmatrix} L_{0,n}(\tau) \\ L_{1,n}(\tau) \\ \vdots \\ L_{K,n}(\tau) \end{pmatrix},$$

and note that $(\alpha_n,\beta_n)^\intercal = \boldsymbol{S}\boldsymbol{L}_n$, where

$$\boldsymbol{S} := (\boldsymbol{X}^\intercal\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Omega} \,. \tag{7}$$

3.3. *Inference*

In this section, we establish an asymptotic distribution for β_n, which serves as the basis for construction of Wald-style confidence intervals and hypothesis tests. We begin by establishing asymptotic linearity of $F^z_{j,n}(\tau)$ for a given j,z. We show that asymptotic linearity immediately implies a joint distribution for $\boldsymbol{F}_n(\tau) := (F^0_{0,n}(\tau),F^1_{0,n}(\tau),\ldots,F^0_{K,n}(\tau),F^1_{K,n}(\tau))^\intercal$, an estimator of $\boldsymbol{F}_0 := (F^0_{0,0}(\tau),F^1_{0,0}(\tau),\ldots,F^0_{K,0}(\tau),F^1_{K,0}(\tau))^\intercal$. We subsequently use this joint distribution and the delta method to derive an asymptotic distribution for β_n.

By definition, an estimator $\hat{F}^z_{j,n}(\tau)$ of $F_{j,0}(\tau)$ is asymptotically linear if $\hat{F}^z_{j,n}(\tau) - F^z_{j,0}(\tau) = \frac{1}{n}\sum_{i=1}^{n} D_0(O_i) + o_P(n^{-1/2})$, where D_0 is a mean-zero, finite-variance function of the observed data that is referred to as the *influence function* of $\hat{F}^z_{j,n}(\tau)$ (Hampel, 1974). The central limit theorem implies that $n^{1/2}\{\hat{F}^z_{j,n}(\tau) - F^z_{j,0}(\tau)\}$ converges in distribution to a normally distributed random variable with mean zero and variance $E_0\{D_0(O)^2\}$. In previous work, we have shown that $F^z_{j,n}(\tau)$ is asymptotically linear and have derived its influence function (Benkeser et al., 2018). We restate those results here. Define $\zeta^z_0(w) := P_0(Z = z \mid W = w)$ as the conditional probability of vaccine and for $t = 1,\ldots,\tau-1$ define $\pi^z_0(t)(w) := P_0(C(t) = 0 \mid Z = z, C(t-1) = 0, N(t) = 0, W = w)$ as the conditional hazard of censoring. Define

$$A^z_0(t)(o) := \frac{I(z = 1, n_\cdot(t-1) = 0, c(t-1) = 0)}{\zeta^z_0(w)\prod_{s=1}^{t-1}\pi^z_0(s)(w)} \,,$$

$$B^z_{j,0}(t)(w) := \sum_{s=t+1}^{t_0}\left[\lambda^z_{j,0}(s)(w)\prod_{m=t+1}^{s-1}\left\{1-\lambda^z_{\cdot,0}(m)(w)\right\}\right] .$$

 ISSN: 2102-6238

The influence function of $F^z_{j,n}(\tau)$ is

$$D^z_{j,0}(o) := \sum_{t=1}^{\tau} \Bigg(A^z_0(t)(o)\{1 - B^z_{j,0}(t)(w)\}\{n_j(t) - \lambda^z_{j,0}(t)(w)\} - A^z_0(t)(o) B^z_{j,0}(t)(w) \left[\sum_{\substack{i=1 \\ i \neq j}}^{K} \left\{ n_i(t) - \lambda^z_{i,0}(t)(w) \right\} \right] \Bigg) + B^z_{j,0}(0)(w) - F^z_{j,0}(\tau) \, . \tag{8}$$

Obtaining the joint distribution of several asymptotically linear estimators is straightforward: by the asymptotic linearity of each component estimators of $\boldsymbol{F}_n(\tau)$,

$$n^{1/2}\{\boldsymbol{F}_n(\tau) - \boldsymbol{F}_0(\tau)\} = n^{1/2} \left\{ \begin{pmatrix} F^0_{0,n}(\tau) \\ \vdots \\ F^1_{K,n}(\tau) \end{pmatrix} - \begin{pmatrix} F^0_{0,0}(\tau) \\ \vdots \\ F^1_{K,0}(\tau) \end{pmatrix} \right\} = \begin{pmatrix} \frac{1}{n^{1/2}} \sum_{i=1}^{n} D^0_{0,0}(O_i) + o_P(1) \\ \vdots \\ \frac{1}{n^{1/2}} \sum_{i=1}^{n} D^1_{K,0}(O_i) + o_P(1) \end{pmatrix} . \tag{9}$$

By the multivariate central limit theorem, (9) converges in distribution to a mean-zero multivariate normal distribution with covariance matrix $E_0\{\boldsymbol{D}_0(O)\boldsymbol{D}_0^\intercal(O)\}$ where $\boldsymbol{D}_0 := (D^0_{0,0}, D^1_{0,0}, \ldots, D^0_{K,0}, D^1_{K,0})^\intercal$.

Based on this joint distribution, we can use the delta method to derive a distribution for β_n. We use $\boldsymbol{S}_{[i,j]}$ to denote the (i,j) entry in $\boldsymbol{S}$, as defined in (7), and note that $\beta_0 = h(\boldsymbol{F}_0)$, where

$$h(\boldsymbol{F}_0) := \sum_{j=0}^{K} \boldsymbol{S}_{[2,j+1]} \log\left\{ \frac{F^0_{j,0}(\tau)}{F^1_{j,0}(\tau)} \right\} .$$

The delta method implies that $n^{1/2}(\beta_n - \beta_0)$ converges in distribution to a mean-zero normally distributed variate with variance

$$\sigma_0^2 := \nabla h(\boldsymbol{F}_0)^\intercal E_0\{\boldsymbol{D}_0(O)\boldsymbol{D}_0^\intercal(O)\}\nabla h(\boldsymbol{F}_0) \, , \tag{10}$$

where ∇h is the gradient of h,

$$\nabla h(\boldsymbol{F}_0) = \left(\frac{\boldsymbol{S}_{[2,1]}}{F^0_{0,0}(\tau)}, -\frac{\boldsymbol{S}_{[2,1]}}{F^1_{0,0}(\tau)}, \ldots, \frac{\boldsymbol{S}_{[2,K+1]}}{F^0_{K,0}(\tau)}, -\frac{\boldsymbol{S}_{[2,K+1]}}{F^1_{K,0}(\tau)} \right)^\intercal .$$

A consistent estimate of σ_0^2 can be computed by substituting the gradient evaluated at $\boldsymbol{F}_n$ and estimated influence function into (10). Specifically, we define $D^z_{j,n}$ to be the estimated influence function that substitutes $\lambda^z_{j,n}$, $\lambda^z_{\cdot,n}$, and empirical estimates of ζ_0 and $\{\pi^z_0(t) : t\}$ into (8). We additionally define $\boldsymbol{D}_n := (D^0_{0,n}, D^1_{0,n}, \ldots, D^0_{K,n}, D^1_{K,n})^\intercal$ and $\boldsymbol{F}_n := (F^0_{0,n}, F^1_{0,n}, \ldots, F^0_{K,n}, F^1_{K,n})^\intercal$. An estimate of σ_0^2 is $\sigma_n^2 := \nabla h(\boldsymbol{F}_n)^\intercal [n^{-1} \sum_{i=1}^n \{\boldsymbol{D}_n(O_i)\boldsymbol{D}_n^\intercal(O_i)\}] \nabla h(\boldsymbol{F}_n)$. This variance estimate may be used to construct asymptotic $100(1-\alpha)\%$ confidence intervals of the form $\beta_n \pm z_{1-\alpha/2} n^{-1/2} \sigma_n$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Similarly, two-sided level α Wald-style hypothesis tests of the null hypothesis that $\beta_0 = 0$ may be performed by rejecting the null hypothesis whenever $|n^{1/2}\beta_n/\sigma_n| > z_{1-\alpha/2}$.

Remark: While not the focus of the present work, if W features many discrete components, then the empirical estimator will likely have high variance and perform poorly in finite samples. If W contains continuous-valued variates, then the stratified estimator cannot be constructed without some discretization of covariates, which risks incurring bias. In these cases, rather than empirical hazard estimators, we may prefer instead regression-based hazard functions that borrow information over time and across covariate levels. The bias-variance tradeoff for the hazard regression may be optimized through the use of cross-validated estimator selection or by regression stacking, also known as super learning (Wolpert, 1992; Breiman, 1996; van der Laan et al., 2007). This approach allows for pre-specification of a large

 ISSN: 2102-6238

number of candidate regressions, which can include parametric regression as well as adaptive regression approaches. Cross-validation is used to estimate the convex combination of candidate regression estimators that optimizes a user-selected risk criteria. Under assumptions, the selected combination of regression functions estimates the true hazard function essentially as well as the unknown best combination of regression function estimates (van der Laan et al., 2004). Stitelman et al. (2011) discusses super learning in the context of estimating hazard functions. We note, however, that if these techniques are used instead of empirical estimators in (6), $\boldsymbol{F}_n$ is no longer asymptotically linear and performing valid inference for estimates of β_0 is challenging. Techniques from the semiparametric efficiency theory literature, such as targeted minimum loss-based estimation (van der Laan and Rubin, 2006), may be used to modify initial hazard estimates in such a way that the estimator (6) based on the modified hazard estimators is asymptotically linear. Benkeser et al. (2018) discusses targeted minimum loss-based estimation in this context. These methods are implemented in the R package survtmle freely available from the Comprehensive R Archive Network (Benkeser and Hejazi, 2017).

3.4. *Choice of weights*

We now turn to selection of the weight matrix $\boldsymbol{\Omega}$. At many amino acid positions, a specific residue is required for biological viability of the pathogen, and, at positions that tolerate multiple residues, viable residues typically have different associations with pathogen fitness. Because of these physiological constraints, endpoints with certain genetic distances may be uncommon. Therefore, we may wish to give more weight to genetic distances that are commonly observed in the population of interest, for example, by weighting estimates of $L_{j,0}$ proportional to the inverse asymptotic covariance matrix of an efficient estimator of $\boldsymbol{L}_0 := (L_{0,0}, \ldots, L_{K,0})^\intercal$. If this matrix were known, the efficiency bound for estimation of projection parameter defined by this choice of weights would have the lowest efficiency bound of any choice of weight matrix (Aitken, 1936).

In practice, the covariance matrix of $\boldsymbol{L}_0$ is unknown, but can be estimated using the influence function methodology of the previous section. We define $g(\boldsymbol{F}_0) := (\log\{F^0_{0,0}/F^1_{0,0}\}, \ \ldots \ , \log\{F^0_{K,0}/F^1_{K,0}\})^\intercal$ and the Jacobian of g as

$$\nabla g(\boldsymbol{F}_0) := \begin{pmatrix} \frac{1}{F^0_{0,0}(\tau)} & -\frac{1}{F^1_{0,0}(\tau)} & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \frac{1}{F^0_{1,0}(\tau)} & -\frac{1}{F^1_{1,0}(\tau)} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \frac{1}{F^0_{K,0}(\tau)} & -\frac{1}{F^1_{K,0}(\tau)} \end{pmatrix} .$$

An estimate of the covariance matrix of $\boldsymbol{L}_0$ is $\boldsymbol{\Upsilon}_n := \nabla g(\boldsymbol{F}_n)[n^{-1}\sum_{i=1}^n \{\boldsymbol{D}_n(O_i)\boldsymbol{D}_n(O_i)^\intercal\}]\nabla g(\boldsymbol{F}_n)^\intercal$. The projection parameter of interest based on this choice of weights is $(\alpha_{0n}, \beta_{0n})^\intercal = (\boldsymbol{X}^\intercal\boldsymbol{\Upsilon}_n^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Upsilon}_n^{-1}\boldsymbol{L}_0$. We index the parameter by n to denote that it depends on the sample. Such parameters are sometimes referred to as data-adaptive, in that they are unknown until one has seen the data (Hubbard et al., 2016). Similarly as above, an estimate of this parameter is

$$(\alpha_n, \beta_n)^\intercal = (\boldsymbol{X}^\intercal\boldsymbol{\Upsilon}_n^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Upsilon}_n^{-1}\boldsymbol{L}_n \, . \tag{11}$$

Theorem 3 of Hubbard et al. (2016) establishes conditions for asymptotic linearity of estimators of data-adaptive target parameters. Under the conditions of this theorem, we may apply our previous results without modification for the estimated weight matrix.

An alternative approach is to define the target parameter as a weighted L^2 projection based on $\boldsymbol{\Upsilon}_0$, the true asymptotic covariance matrix of an efficient estimator of $\boldsymbol{L}_0$. The inference we have derived would likely be anti-conservative for estimation of this parameter as we ignore uncertainty induced by estimation of $\boldsymbol{\Upsilon}_0$. However, a nonparametric bootstrap wherein in each resample one estimates both $\boldsymbol{\Upsilon}_0$ and the projection parameter may well lead to confidence intervals with proper coverage. We leave this study to future work.

 ISSN: 2102-6238

4. Simulation

We studied the performance of our estimators of β_{0n} via simulation. A single covariate W was drawn from a Binomial$(4, 0.5)$ distribution, which mimics the geographic site variable used in the RTS,S analysis. Vaccine assignment Z was drawn from a Bernoulli(0.5) distribution, which mimics a randomized trial with equal vaccine allocation. Given $W = w$ and $Z = z$ an endpoint time was generated from a geometric distribution with failure probability expit$[-2+0.4\{I(w=0)+I(w=1)+I(w=2)\}-0.2I(w=3)-z]$. Similarly, a censoring time was generated from a geometric distribution with failure probability expit$\{-3+0.2I(w=2)-0.2I(w=3)\}$. The observed failure time was taken to be the minimum of the endpoint and censoring times, with ties recorded as endpoints. Given $Z = z$, the genetic distance associated with each endpoint was drawn from a Binomial$(4, \text{expit}(0.2z))$ distribution. This choice resulted in vaccine efficacy that decreased with genetic distance, the direction most commonly seen in vaccine sieve analysis. We set the final observation at $\tau = 6$, and any observations with an endpoint beyond this time were right-censored at τ. We analyzed 1,000 data sets of size 1000, 2500, and 5000.

Figure 1 shows the true efficacy across genetic distance for this data generating mechanism. The log-ratio of cumulative incidences is greater than zero for each distance indicating efficacy against each distance value, though the efficacy decreases with distance. To illustrate the impact of the estimated weight matrix on the parameter of interest, the figure shows the projection lines with the largest and smallest true value of the slope parameter β_{0n} across all simulations. The two lines are quite similar indicating that the estimated covariance matrix $\boldsymbol{\Upsilon}_n$ was relatively stable and had only a minor effect on the true value of the parameter of interest across simulations.

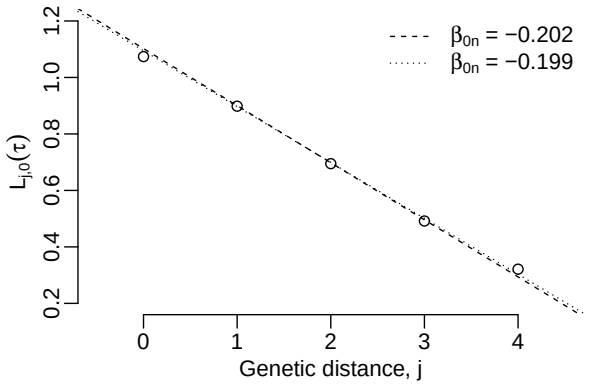

FIGURE 1. *True values of the log ratio of cumulative incidences by genetic distance (circles). The two lines are the lines associated with the most extreme true slope parameters across all simulated data sets.*

For each of the sample sizes considered, the estimators of β_{0n} were approximately unbiased (Table 1). The bias and variance of the estimators decreased appropriately with sample size and the confidence intervals achieved approximately nominal coverage at all sample sizes. Overall, the estimators had excellent finite-sample performance.

	Bias x 1e2	Variance x 1e2	MSE x 1e2	Coverage %
$n = 1,000$	0.59	1.20	1.20	94.2
$n = 2,500$	0.38	0.45	0.46	94.9
$n = 5,000$	0.28	0.21	0.21	96.1

TABLE 1. *Results from the simulation study showing bias, variance, and mean squared-error (MSE) of β_n as an estimator of β_{0n}. The coverage of nominal 95% Wald-style confidence intervals for β_{0n} is also shown.*

 ISSN: 2102-6238

5. Data analysis

The design of the Phase III RTS,S/AS01 trial has been previously described (RTS,S Clinical Trials Partnership, 2011; Agnandji et al., 2012), as have the methods for *Plasmodium* parasite sequencing (Neafsey et al., 2015). We illustrate the proposed method by studying the efficacy of the RTS,S efficacy as a function of Hamming distance, i.e., the number of mismatched amino acid residues, between an aligned founder sequence and the vaccine insert sequence in the Th3R epitope region contained within the sequenced C-terminus amplicon of the CS protein. Th3R is a putative T-cell epitope region that is 12 amino acids long. Figure 2 shows the distribution of the Th3R distance from the RTS,S vaccine by treatment arm in children aged 5 to 17 months. We also analyzed Hamming distances defined based on the entire C-terminus amplicon of CS and on three other pre-specified epitope regions, but for brevity restrict reporting of results to the Th3R distances.

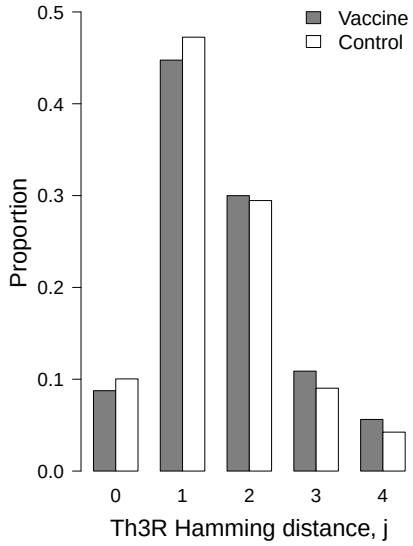

FIGURE 2. *Distribution of the Th3R Hamming distance of clinical malaria sequences to the vaccine insert sequence in children aged 5–17 months*

A unique feature of the pathogen sequence collection was that multiple founder parasites could be sequenced from the dried blood spot sample of a single participant. The majority of the 2,090 clinical malaria endpoints through $\tau = 12$ months post-vaccination with available C-terminus sequence data were found to have multiple parasite genotypes, with each founder parasite likely caused by a transmission event from a distinct mosquito. The presence of multiple founder infections complicates the assessment of sieve effects by genetic distance, as a single participant may have several founder parasites each with a unique genetic distance. We used multiple outputation (Follmann et al., 2003) to estimate the trend in the vaccine's effect across the Th3R Hamming distance for a *randomly sampled* founder parasite of a clinical malaria case. This approach separates the genetic component of a sieve effect from any effect that the vaccine might have on the number of infecting parasites (Neafsey et al., 2015). Multiple outputation was performed by repeatedly sampling a single parasite genotype at random from each clinical malaria endpoint and applying a statistical method designed for a single genotype per endpoint. Based on the guidelines of Follmann et al. (2003), the estimated number of outputations required for stable inference was $B = 4,127$. For each resample, we estimated $L_{j,0}(\tau)$, $j = 0,\ldots,4$, adjusting for geographical study site as the sole covariate, and averaged of these estimates over resamples. We then estimated and tested the proposed trend parameter on the outputation-averaged estimates of $L_{j,0}(\tau)$ weighting by the estimated inverse variance of the multiple-outputation estimates.

Point estimates of $\mathrm{VE}_{j,0}(\tau)$ with 95% Wald-style confidence intervals (CIs) are plotted in Figure 3, with $\mathrm{VE}_{0,n}(\tau) = 46\%$ (95% CI, 32 to 57%) against a perfectly vaccine-matched parasite in the Th3R region and a steady decline to $\mathrm{VE}_{4,n}(\tau) = 11\%$ (95% CI, -25 to 36%) against parasites with four

 ISSN: 2102-6238

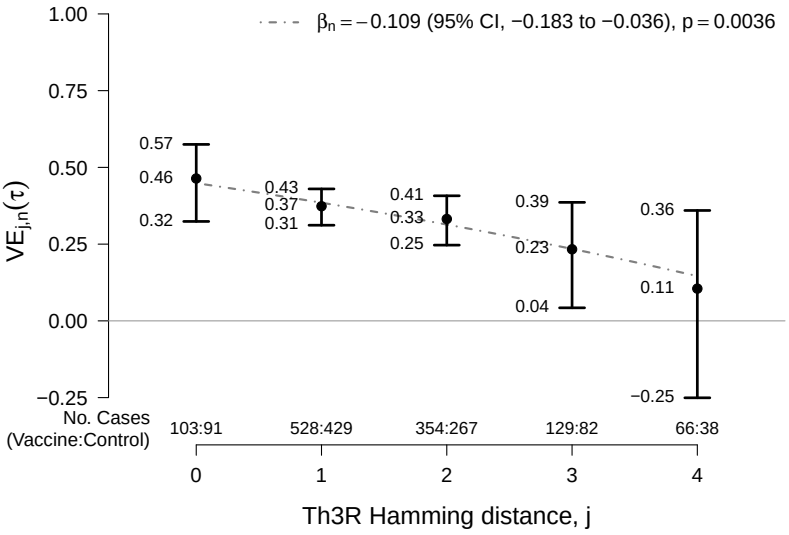

FIGURE 3. *Point estimates of $VE_{j,0}(\tau)$ against clinical malaria with Th3R Hamming distance $j = 0,\ldots,4$ to the vaccine insert sequence, with 95% Wald-style confidence intervals, in children aged 5–17 months. The superimposed curve $1 - exp\{-\alpha_n - \beta_n j\}$ is a transformation to the VE scale of the linear projection of $L_{j,n}(\tau)$. The two-sided Wald-style test of the null hypothesis that $\beta_0 = 0$ yields the p-value of 0.0036.*

vaccine-mismatched Th3R residues. The two-sided Wald-style test of the null hypothesis that $\beta_0 = 0$ yields the p-value of 0.0036 suggesting potential immunological relevance of Th3R epitopes for multivalent vaccine design.

6. Discussion

The proposed trend parameters are appealing in the context of vaccine sieve analysis in that we often expect trends in efficacy by genetic distance to be monotone. Thus, the slope of the projection onto a linear working model provides a reasonable summary of the trend. In other applications, monotonicity may not be expected. For example, in cardiovascular epidemiology, researchers are interested different types of heart failure, which are defined by the strength of the heart's contraction using a measure called ejection fraction. Our methods could be applied to estimate the trend in a relationship between biomarkers and ejection fraction. However, we might not have a reason to expect monotonicity in this relationship. Nevertheless, our methods easily extend to more flexible working models (e.g., including polynomial terms) with simple modifications to the delta method calculus. Multiple-degree-of-freedom tests could be developed to test whether there is any variation in the effect of a treatment across levels of an ordinal competing risk.

The efficacy of vaccines often wanes over time, due to waning immune responses many months or years after receipt of vaccinations, and this occurred in the Phase III RTS,S efficacy trial. While our method makes no assumption of time constancy of the vaccine's effect, it does require the selection of a single fixed time at which to examine trends in efficacy by pathogen genetic distance, and these trends may vary over time. Therefore, an interesting elaboration of our method would extend the working models to summarize trends in efficacy across distance and over time. Working models could include functions of both genetic distance and time, while hypothesis tests could test how the trend in efficacy varies over time. Another interesting extension would provide simultaneous confidence bands or regions for the pattern of vaccine efficacy as it varies over genetic distance and/or time.

A code repository, which can be downloaded as an R package, is available (https://github.com/benkeser/sievetrend). The repository contains the code used to execute the simulation study and data analysis.

 ISSN: 2102-6238

Acknowledgements

The authors thank the RTS,S/AS01 Phase 3 trial study participants and study investigators, and in particular thank Daniel Neafsey at the Department of Immunology Broad Institute of Massachusetts Institute of Technology and Dyann Wirth at the Harvard T.H. Chan School of Public Health for collaboration and generation of the malaria parasite sequence data.

References

Agnandji, S., Lell, B., Fernandes, J., Abossolo, B., Methogo, B., Kabwende, A., Adegnika, A., Mordmüller, B., Issifou, S., et al. (2012). A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants. *New England Journal of Medicine*, 367(24):2284–95.

Aitken, A. C. (1936). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48.

Benkeser, D., Carone, M., and Gilbert, P. B. (2018). Improved estimation of the cumulative incidence of rare outcomes. *Statistics in Medicine*, 37(2):280–293.

Benkeser, D. and Hejazi, N. S. (2017). *survtmle: Targeted Minimum Loss-Based Estimation for Survival Analysis in R*.

Breiman, L. (1996). Stacked regressions. *Machine learning*, 24(1):49–64.

Follmann, D., Proschan, M., and Leifer, E. (2003). Multiple outputation: Inference for complex clustered data. *Biometrics*, 59:420–429.

Gilbert, P. B., McKeague, I. W., and Sun, Y. (2008). The two-sample problem for failure rates depending on a continuous mark: An application to vaccine efficacy. *Biostatistics*, 9(2):263–276.

Gilbert, P. B., Self, S. G., and Ashby, M. A. (1998). Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics*, 54(3):799–814.

Gilbert, P. B., Self, S. G., Rao, M., Naficy, A., and Clemens, J. (2001). Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *Journal of Clinical Epidemiology*, 54(1):68–85.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hubbard, A. E., Kherad-Pajouh, S., and van der Laan, M. J. (2016). Statistical inference for data adaptive target parameters. *The International Journal of Biostatistics*, 12(1):3–19.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Juraska, M. and Gilbert, P. B. (2013). Mark-specific hazard ratio model with multivariate continuous marks: An application to vaccine efficacy. *Biometrics*, 69(2):328–337.

Juraska, M. and Gilbert, P. B. (2016). Mark-specific hazard ratio model with missing multivariate marks. *Lifetime Data Analysis*, 22(4):606–625.

Neafsey, D. E., Juraska, M., Bedford, T., Benkeser, D., Valim, C., Griggs, A., Lievens, M., Abdulla, S., Adjei, S., Agbenyega, T., et al. (2015). Genetic diversity and protective efficacy of the RTS,S/AS01 malaria vaccine. *New England Journal of Medicine*, 373(21):2025–2037.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and van der Laan, M. J. (2010). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54.

Rolland, M., Edlefsen, P. T., Larsen, B. B., Tovanabutra, S., Sanders-Buell, E., Hertz, T., Carrico, C., Menis, S., Magaret, C. A., and Ahmed, H. (2012). Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature*, 490(7420):417–420.

RTS,S Clinical Trials Partnership (2011). First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *New England Journal of Medicine*, 365(20):1863–1875. PMID: 22007715.

Stitelman, O. M., Wester, C. W., De Gruttola, V., and van der Laan, M. J. (2011). Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *The International Journal of Biostatistics*, 7(1):1–34.

Sun, Y., Gilbert, P. B., and McKeague, I. W. (2009). Proportional hazards models with continuous marks. *Annals of Statistics*, 37(1):394.

Sun, Y., Li, M., and Gilbert, P. B. (2012). Mark-specific proportional hazards model with multivariate continuous marks and its application to HIV vaccine efficacy trials. *Biostatistics*, 14(1):60–74.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):1–40.

van der Laan, M. J., Dudoit, S., and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):1–23.

WHO (2016). Weekly epidemiological record. *World Health Organization*, 91:33–52.

Wolpert, D. (1992). Stacked generatlization. *Neural Networks*, 5(2):241–259.

 ISSN: 2102-6238