Published in final edited form as:

Biometrics. 2009 December ; 65(4): 1223–1232. doi:10.1111/j.1541-0420.2009.01189.x.

Assessing Vaccine Effects in Repeated Low-Dose Challenge Experiments

Michael G. Hudgens^{1,*} and Peter B. Gilbert²

¹ Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A

² Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, Washington, 98109, U.S.A

SUMMARY

Evaluation of HIV vaccine candidates in non-human primates (NHPs) is a critical step toward developing a successful vaccine to control the HIV pandemic. Historically, HIV vaccine regimens have been tested in NHPs by administering a single high dose of the challenge virus. More recently, evaluation of candidate HIV vaccines has entailed repeated low-dose challenges which more closely mimic typical exposure in natural transmission settings. In this paper, we consider evaluation of the type and magnitude of vaccine efficacy from such experiments. Based on the principal stratification framework, we also address evaluation of potential immunological surrogate endpoints for infection.

Keywords

Causal inference; Correlates of protection; HIV; Potential outcomes; Surrogate marker; Vaccine trial

1 Introduction

As of 2007, approximately 33.2 million people were living infected with HIV, with over 2.1 million people dying of AIDS in that year (UNAIDS 2007). While great strides have been made in developing effective antiretroviral therapy for treatment of HIV infected individuals, a preventive vaccine remains the greatest hope in curbing the HIV pandemic. HIV vaccine research begins with in vitro and animal studies. A critical component of this pre-clinical development entails evaluation of candidate vaccines in non-human primates (NHPs) such as macaques. Historically, HIV vaccine regimens have been tested in NHPs by administering a single high dose of the challenge virus. More recently, evaluation of candidate HIV vaccines has entailed repeated low-dose (RLD) challenges which more closely mimic typical exposure to HIV in natural transmission settings (Regoes et al. 2005; Subbarao et al. 2006; Ellenberger et al. 2006). A primary objective of these studies is to assess vaccine efficacy for prevention of infection. A secondary objective is to determine immune biomarkers which are surrogate endpoints for infection, which we refer to as “surrogates of protection.”

Since the RLD challenge study design has only recently been implemented in evaluation of candidate HIV vaccines, the corresponding statistical literature is rather limited. One exception is Regoes et al. (2005), who show that for clinically feasible samples sizes, RLD challenge

*e-mail: mhudgens@bios.unc.edu.

Supplementary Materials

The Web Appendix, Tables, and Figure referenced in Section 2 and 3 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

studies in NHPs can be adequately powered to test for vaccine efficacy to prevent infection. This rather surprising result is due to the exquisitely precise nature of the exposure and infection history in challenge studies. In contrast, studies of HIV in humans typically provide only vague information on the number of exposures prior to infection. Consequently RLD challenge studies with small sample sizes can have the same power as large phase III clinical trials to test for vaccine efficacy.

Beyond testing for a vaccine effect, it is not clear what additional information can be inferred from RLD challenge studies. For example, can these studies inform about the type or magnitude of a vaccine's protective effect? Accurately characterizing the mechanism of protection would provide important information for the design and analysis of future efficacy trials and for population models on the impact of a licensed vaccine. Additionally, is it possible to evaluate potential immune surrogates of protection in RLD challenge studies? This paper seeks to answer these questions. In Section 2 we consider evaluating the type and magnitude of vaccine efficacy from RLD challenge experiments; an illustrative example is given using recently published results from a challenge study of a candidate HIV vaccine. In Section 3 we describe a causal inference approach to assessing potential immunological surrogates of protection in this setting. We conclude with a discussion in Section 4. Similar to Regoes et al. (2005), we find that, despite relatively small sample sizes, RLD challenge studies can provide accurate and precise information about vaccine efficacy and immune surrogate markers. While this work is motivated by the development of an HIV vaccine, the proposed methods can easily be applied to RLD challenge studies of other vaccines and other preventive interventions.

2 Vaccine efficacy for susceptibility

In this section we consider evaluating the type and magnitude of vaccine efficacy for susceptibility (VE_s), i.e., a vaccine's ability to protect against infection. Our approach entails applying maximum likelihood methods to a discrete time survival model which allows for possible heterogeneous vaccine effects (Halloran et al. 1992; Longini and Halloran 1996).

2.1 Discrete time survival model

Let p denote the probability of transmission to a susceptible, unvaccinated individual given a single exposure (i.e., challenge). Assuming the probability of infection is independent of the number of prior challenges, the probability of escaping infection from t challenges for an unvaccinated individual is $(1 - p)^t$. On the other hand, the probability an unvaccinated individual becomes infected on the t^{th} challenge is $(1 - p)^{t-1}p$. Next suppose a vaccine may confer partial or complete protection to individuals. Let θ be the probability a vaccinated individual is completely protected. Further, suppose the probability a vaccinated individual who is not completely protected gets infected from a single challenge is ϕp . Then the probability a vaccinated individual escapes infection from t challenges is $(1 - \phi p)^t(1 - \theta) + \theta$, whereas the probability a vaccinated individual becomes infected from challenge t is $(1 - \phi p)^{t-1}\phi p(1 - \theta)$.

Since the probability a vaccinated individual becomes infected from a single challenge is $(1 - \theta)\phi p$, define the vaccine efficacy as

$$VE_s \equiv 1 - \frac{(1 - \theta)\phi p}{p} = 1 - (1 - \theta)\phi, \quad (1)$$

i.e., the relative reduction in the per contact transmission probability if vaccinated compared to if not vaccinated. This measure of vaccine efficacy has been referred to as the *per contact* or *biological* efficacy (Halloran et al. 1999). If $\theta > 0$ and $\phi = 1$, then each vaccinated individual

is either completely protected or not at all protected. In this case the vaccine is said to have an “all-or-none” effect with $VE_S = \theta$. On the other hand, if $\theta = 0$ and $\varphi < 1$, then only partial protection is conferred to all vaccinees whereby each vaccinated individual has a reduced transmission probability by the same multiplicative factor. In this case the vaccine is said to have a “leaky” effect with $VE_S = 1 - \varphi$. If $\theta > 0$ and $\varphi \neq 1$, the vaccine has a “mixed” effect with both all-or-none and leaky mechanisms of protection with efficacy (1). Finally, if $\theta = 0$ and $\varphi = 1$, then the vaccine has no efficacy, i.e., $VE_S = 0$; we refer to this situation as the “null” model.

Maximum likelihood methods (see Web Appendix A) can be employed for inference regarding (p, θ, φ) based on data from a NHP challenge study such as described below in Section 2.2. For such a study, we consider point and interval estimation of VE_S , as well as model selection from among the four mechanism of protection models described above. While it is difficult to discern a vaccine’s protective mechanism from a large human vaccine efficacy trial (Farrington 1998; Gilbert 2001), it is more feasible in an RLD challenge study, due to the far greater information on exposure and transmission.

2.2 Example

Ellenberger et al. (2006) employed a RLD challenge study in macaques to assess the efficacy of a candidate HIV vaccine. The vaccine was given to 16 macaques and 14 additional macaques served as controls, i.e., did not receive the vaccine. All animals were then repeatedly exposed weekly to a hybrid simian-human immunodeficiency virus (SHIV) with a different HIV sequence than the HIV sequence represented in the vaccine. Evidence of systemic infection was assessed after each exposure. Infection was defined as having detectable cell-free virus and provirus in peripheral blood mononuclear cells. It is assumed that the cell-free and cell-associated diagnostic tests used for infection diagnosis were sufficiently accurate and the weekly time intervals between challenges were far enough apart such that determination of the infecting exposure was made without error. Four monkeys in the vaccine arm and one in the control arm were administratively right censored after escaping infection from multiple exposures.

Data from this experiment are given in Web Table 1 and the corresponding maximum likelihood results are given in Table 1. Based on a likelihood ratio test (LRT) comparing the leaky and null models, there is evidence of a significant leaky vaccine effect (p-value=0.003). A one-sided Fisher’s exact test (Regoes et al. 2005) gives a similar result (p-value=0.006). Comparison of the log likelihood values for the leaky and all-or-none models suggests the leaky model demonstrates superior fit. Similarly, the LRT comparing the mixed and leaky models is not significant (p-value=0.2). The Akaike Information Criterion (AIC) also suggests the leaky model provides the most parsimonious model that adequately fits the given data.

The Kaplan-Meier estimates of the survival function for each arm of the study are given in the left panel of Figure 1. The agreement between these nonparametric estimates and the corresponding estimates from the leaky vaccine model evaluated at the MLEs $\hat{\varphi} = 0.36$ and $\hat{p} = 0.20$ suggest good fit of this model to the data. The nonparametric estimates of the complementary log-log survival curves in the right panel of Figure 1 are roughly parallel, also supporting a leaky vaccine effect (Halloran et al. 1999). To evaluate the mechanism of protection further, simulated data sets were generated from the four different models under consideration (evaluated at the MLEs from Table 1) to provide a basis of comparison. Figure 2 shows the difference in the non-parametric estimates of the complementary log-log survival curves from the observed data and 25 simulated data sets from the four models. These plots also suggest the leaky model provides a better fit than the all-or-none or null models.

To formally test for goodness-of-fit, the following Kolmogorov-Smirnov type test statistic was computed:

$$T_{KS} = \max\{\sup_t |S_0^c(t) - \hat{S}^c(t)|, \sup_t |S_0^v(t) - \hat{S}^v(t)|\}$$

where $S_0^c(t)$ and $S_0^v(t)$ denote the survival curves for controls and vaccinees under the leaky model when $p = 0.20$ and $\varphi = 0.36$, and $\hat{S}^c(t)$ and $\hat{S}^v(t)$ are the corresponding nonparametric estimates. The sampling distribution of T_{KS} was approximated by simulating 10,000 data sets under the null hypothesis $p = 0.20$, $\varphi = 0.36$, and $\theta = 0$. The resulting approximate p-value was 0.35, indicating adequate fit of the leaky model.

In total, these results suggest a significant leaky vaccine effect. The MLE of VE_S is 0.64, which can be interpreted as a 64% reduction in the probability of infection per exposure. The Jewell (1986) (see also Chick et al. 2001) bias corrected estimate of VE_S is 0.66. The profile likelihood 95% confidence interval (CI) for VE_S is (0.26, 0.83).

2.3 Simulation studies

Simulation studies were conducted to investigate the operating characteristics of several of the statistical methods employed in the example above. Unless stated otherwise, data were simulated assuming a transmission probability under control of $p = 0.2$, an equal number of NHPs in the vaccine and control arms, and the experiment ceases after 20 exposures if a NHP is still SHIV negative.

The first set of simulations assumed a leaky mechanism of protection, i.e., $\theta = 0$. Based on 10,000 simulations, the power of the LRT comparing the leaky and null models to detect a departure from the null hypothesis $H_0: \varphi = 1$ at the $\alpha = 0.1$ level is given in the top half of Table 2. For purposes of comparison with the power to detect an all-or-none effect (discussed below), here we assumed $\varphi \leq 1$, i.e., the vaccine does not increase the probability of infection per exposure in individuals who are not completely protected. Consequently, since φ is on the boundary of the parameter space under H_0 , the distribution of the LRT statistic was assumed to be $0.5\chi_0^2 + 0.5\chi_1^2$, i.e., a 50:50 mixture of chi-squared distributions with 0 and 1 degrees of freedom (Self and Liang 1987). The results of Table 2 suggest the LRT preserves the type I error for sample sizes typical of challenge studies and that the Ellenberger et al. (2006) experiment had over 80% power to detect a leaky VE_S of 60%. Similar results were found using a Wald test (results not shown). These findings are in agreement with Regoes et al. (2005), who also showed low-dose challenge studies can be adequately powered to detect leaky vaccine effects.

The second set of simulations assumed an all-or-none mechanism of protection, i.e., $\varphi = 1$. The simulation results in the lower portion of Table 2 give the power for comparing the all-or-none and null models to detect a departure from the null $H_0: \theta = 0$. Again the distribution of the LRT statistic was assumed to be $0.5\chi_0^2 + 0.5\chi_1^2$. Comparison of the upper and lower portions of Table 2 indicates there is greater power to detect an all-or-none effect than a leaky effect. Additional simulations (results not shown) indicate the LRT also preserves the type I error rate when comparing the mixture and leaky models.

A simulation study was also employed to assess the ability of the AIC to select the correct mechanism of protection model (i.e., null, all-or-none, leaky, or mixed). The results in Web Table 2 give the probability of selecting the correct model for different values of p , θ , φ , and the maximum number of allowable challenges C^{max} . If the true model is all-or-none or null, the probability of selection is typically adequate, i.e., approximately 0.9. Correct selection can

be substantially less likely for leaky or mixture models. For example, if $p = 0.2$ and $C^{max} = 30$, there is only a 0.01 probability of correctly selecting a mixture model when $\theta = 0.8$ and $\varphi = 0.2$. However, by increasing p to 0.5 (e.g., by increasing the challenge dose) or increasing C^{max} , the probability of correct selection for leaky and mixture models increases. Nonetheless, these results indicate model selection should not be based on AIC alone in this setting.

Finally, we also examined the bias of the MLE of VE_S under the leaky model. Simulation results in Web Table 3 demonstrate an appreciable negative bias of $\widehat{VE}_S$ for smaller sample sizes. The Jewell bias corrected estimator (Jewell 1986; Chick et al. 2001) appears to be preferable in terms of bias; Chick et al. (2001) reached a similar conclusion in the context of vaccine efficacy evaluation in small or intermediate size (e.g., Phase IIb) clinical trials. For each simulated data set, we also computed a profile likelihood 95% CI for VE_S . The empirical coverage probability of these CIs given in Web Table 3 is very close to the nominal coverage probability.

2.4 Elaborations

2.4.1 Accommodating heterogeneous transmission probabilities—To this point we have assumed a homogeneous transmission probability p , i.e., that every individual has the same natural susceptibility to infection at each time point. This assumption may be violated due to among-individual variability in host genetics (e.g., HLA type), immunity, and other characteristics. Failing to account for susceptibility heterogeneity may lead to biased estimation of VE_S and to undercoverage of confidence intervals (Halloran et al. 1992). To relax this homogeneity assumption, we suppose the transmission probabilities vary between NHPs according to a beta distribution. For the control group, the beta distribution of the transmission probability p can be specified by the mean μ and coefficient of variation η . For the vaccine subgroup not completely protected, we suppose the beta distribution of the transmission probability, say p_v , has mean $\varphi\mu$ and coefficient of variation η . That is, the beta distributions for the control and vaccine transmission probabilities are assumed to have the same coefficient of variation but possibly different means. The vaccine efficacy estimand VE_S then takes the same form as in (1), with modified interpretation as the percent change in the expected per-contact transmission probability. Under the additional assumptions that the transmission probabilities are independent across exposures within animals and the individual-specific mean transmission probabilities have a beta distribution across individuals, then the assumption that p and p_v are constant across exposures within individuals is no longer needed (Weinberg and Gladen 1986).

For each of the mixture, leaky, all-or-none, and null models, the MLE of $(\theta, \varphi, \mu, \eta)$, and thus of VE_S , can be computed using the likelihood given in the Web Appendix A. Returning to the RLD challenge study in Section 2.2, maximum likelihood results allowing for heterogeneous transmission probabilities are given in Web Table 4. Comparing these results with Table 1, the AIC still selects a leaky vaccine model with homogeneous transmission probabilities.

2.4.2 A generalized estimand—The definition of vaccine efficacy given in (1) is based on a single contact. More generally, VE_S could be defined in terms of t contacts for $t \geq 1$, i.e.,

$$VE_S(t) = 1 - \frac{(1 - \theta)\{1 - (1 - \varphi p)^t\}}{1 - (1 - p)^t},$$

the relative reduction in the probability of infection from t exposures under vaccine compared to control. If the vaccine is tested or to be used in a low-risk population, then the estimand

$VE_S(1)=VE_S$ might be considered. Conversely if the vaccine is tested in a high-risk population, a definition with more exposures, for example $VE_S(50)$, might be more appropriate. Plots of $VE_S(t)$ given in Web Figure 1 demonstrate the well known result that vaccine efficacy is not affected by the number of exposures for an all-or-none mechanism model, whereas with a leaky or mixed model, the efficacy decreases with an increasing number of exposures (Smith et al. 1984). Web Figure 1 also shows that the greater the transmission probability p , the greater the attenuation effect.

The generalized definition $VE_S(t)$ helps explain the greater observed power (Table 2) for detecting an all-or-none effect compared to a leaky effect. For example, under the leaky model with $\phi = 0.5$, the alternative hypothesis is equivalent to $\{VE_S(t): \theta = 0, \phi = 0.5, t \geq 1\}$, where $VE_S(t)$ is near 0 for t large. In contrast, for the all-or-none model with $\theta = 0.5$, the alternative hypothesis is equivalent to $\{VE_S(t): \theta = 0.5, \phi = 1, t \geq 1\}$, where $VE_S(t) = 0.5$ for all t . Thus the alternative for the all-or-none model can be viewed as farther from the null hypothesis than the leaky model.

3 Surrogates of protection

In this section we present an approach for assessing potential immunological surrogates of protection (SoP) in a RLD challenge study. In general, a SoP is defined to be an immunological variable S such that a vaccine effect on S is predictive of a vaccine effect on the risk of infection or disease (i.e., is predictive of VE_S). The utility of such a surrogate marker includes guiding vaccine development, improving immunogens iteratively between basic and clinical research, providing guidance for regulatory decisions, bridging efficacy of a vaccine observed in a trial to a new setting, and guiding public immunization policy. For RLD challenge studies, knowledge of an immunological surrogate may allow insightful and cost-effective comparisons of vaccine candidates in animals, support predictions of vaccine efficacy in humans, and inform prioritizing the most promising candidates for testing in humans.

Despite the importance of finding SoPs, the literature on methods for their quantitative assessment is quite limited. Most existing approaches simply assess correlates of risk (CoRs), i.e., immunological biomarkers that are associated with risk of infection or disease. For example, in the first phase III trial of an HIV vaccine (VAX004), a significant negative association was found between risk of HIV infection and antibody (Ab) response to the vaccine (Gilbert et al. 2005). However, this purely correlational analysis provides no information to distinguish between the possible explanations that (i) a greater vaccine effect on the immune response predicted a greater vaccine effect on infection risk, or (ii) the immune response simply marked an innate ability to escape infection but did not predict vaccine efficacy. In other words, it was not possible to conclude whether Ab response to the vaccine was a SoP or just a CoR.

Qin et al. (2007) defined a hierarchy of two levels of SoPs: a *specific SoP* is predictive of VE_S for the same setting (population, environmental factors) as present in the particular study, and a *general SoP* is a specific SoP that is also predictive of VE_S across different settings (e.g., across populations or across vaccine formulations). Meta-analysis of multiple vaccine studies is required for evaluating a general SoP whereas one study may be sufficient for evaluating a specific SoP. While developing a general SoP is most valuable scientifically, developing a specific SoP is valuable in itself and as an intermediate step toward developing a general SoP. Moreover, enabling the evaluation of a general SoP is practically challenging, requiring several vaccine studies that provide ample variability in the meta-analytic unit of interest, whereas a specific SoP may be evaluated using a practicable augmentation of existing and planned vaccine studies. Here we restrict attention to the evaluation of a specific SoP from a single RLD challenge study of a candidate HIV vaccine.

Recently, novel experimental designs and corresponding statistical methodology have been proposed for evaluating potential specific SoPs in the context of human efficacy trials (Follmann 2006; Gilbert and Hudgens 2008). Here we consider one of two designs proposed by Follmann (2006) wherein for each individual we measure a baseline covariate(s) W that is correlated with the immune response that individual would have to the HIV vaccine being evaluated. For example, W might be an immune response to a rabies vaccine. The missing HIV vaccine immune response for individuals in the control arm can then be predicted from their W and a prediction model based on observed data from the vaccine group. In turn, we can assess how well causal treatment effects on the HIV immune response predict the causal effect of the vaccine to prevent infection. Simulation studies of large (e.g., Phase III) randomized studies have demonstrated that the additional information provided by W can enable assessing the extent to which a CoR is a SoP (Follmann 2006; Gilbert and Hudgens 2008). Below we consider whether measuring a baseline predictor W in RLD challenge studies might also afford sufficient information for inference regarding possible SoPs. We know of no RLD challenge studies to date which have implemented Follmann's baseline predictor study design.

3.1 Methods

Motivated by Ellenberger et al. (2006), we consider assessment of possible SoPs assuming a leaky vaccine effect. We begin by introducing the potential outcomes notation to be used for the SoP model. For subject i , let $T_i(Z)$ be the potential survival time under assignment to treatment Z for $Z = 0$ (control) or $Z = 1$ (vaccine). We assume throughout that survival time is measured by the number of exposures (i.e., challenges) until infection; thus $T_i(Z)$ is a positive integer. Next let $S_i(Z)$ denote the HIV-specific immune response under treatment Z . We assume $S_i(0) = 0$ for all i since vaccine antigens (absent in the control) must be present to induce an HIV-specific immune response. We also assume $S_i(1) \geq 0$ and that $S_i(1)$ is measured at a single time point after treatment assignment and before the first exposure.

In order to identify the causal estimands of interest defined below, we invoke the stable unit treatment value assumption (SUTVA) and assume ignorable treatment assignment. The lack of interference between NHPs implied by SUTVA should hold in this setting since investigators can prevent interaction between NHPs. Use of randomization in assigning NHPs to receive vaccine or serve as a control will insure ignorable treatment assignment. In the context of human vaccine trials, Gilbert and Hudgens (2008) make the additional assumption that the risk of infection prior to measurement of S_i is the same for $Z = 0$ and $Z = 1$. This additional assumption is not required in RLD challenge studies since investigators control the timing of exposures.

The average causal effect of the vaccine on survival is defined as $h(E\{T_i(0)\}, E\{T_i(1)\})$ where h is some contrast function such that $h(x, y) = 0$ iff $x = y$, e.g., $h(x, y) = x - y$. Next let $p(Z)$ denote the probability of infection from a single exposure (i.e., the transmission probability) under assignment to treatment Z . Assuming the probability of infection is independent of the number of prior exposures, $T_i(Z)$ has a geometric distribution with mean $E\{T_i(Z)\} = 1/p(Z)$ for $Z = 0, 1$, indicating the average causal effect can equivalently be described by a contrast in transmission probabilities. For example, if $h(x, y) = 1 - x/y$, then the causal effect of the vaccine on survival equals $VE_s^{CE} \equiv 1 - p(1)/p(0)$, which is equivalent to (1) (i.e., $VE_s^{CE} = VE_s$) in the absence of an all-or-none vaccine effect.

Next consider the principal stratification (Frangakis and Rubin 2002) of individuals according to the pair of potential immune responses $(S_i(0), S_i(1))$. Since $S_i(0) = 0$ for all i , membership within a principal stratum is determined completely by $S_i(1)$. Following Gilbert and Hudgens (2008), we define S to be an SoP if

$$h(E\{T_i(0)|S_i(1)=0\}, E\{T_i(1)|S_i(1)=0\})=0$$

and

$$h(E\{T_i(0)|S_i(1)=s\}, E\{T_i(1)|S_i(1)=s\}) \neq 0 \text{ for all } s > C,$$

for some constant $C \geq 0$. In words, S is an SoP if the vaccine has no average effect on survival in groups of individuals who would have no immune response under vaccine and has some average effect in groups of individuals who would have an immune response greater than C under vaccine. Let $p(z, s)$ denote the transmission probability conditional on $S_i(1) = s$ and $Z = z$, such that $E\{T_i(z)|S_i(1) = s\} = 1/p(z, s)$. Then S is an SoP if

$$p(0, 0) = p(1, 0) \text{ and } p(0, s) \neq p(1, s) \text{ for all } s > C. \quad (2)$$

Therefore, whether S is an SoP can be evaluated through inference about the transmission probability curves $p(0, s)$ and $p(1, s)$.

In practice, a biomarker may have value as a surrogate even if (2) is not strictly satisfied. For example, if $p(0, 0)$ is approximately equal to $p(1, 0)$ while $p(0, s)$ is substantially greater than $p(1, s)$ for $s > C$, then S is predictive of the vaccine's effect on risk of infection. Therefore, to summarize the predictiveness or "surrogate value" of a biomarker, Gilbert and Hudgens (2008) proposed the proportion associative effect statistic $PAE \equiv |EAE|/(|EAE| + |EDE|)$ where

$$EAE \equiv E[h(E\{T_i(0)|S_i(1)\}, E\{T_i(1)|S_i(1)\})|S_i(1) > S_i(0)] \\ \int_{s>0} h(E\{T_i(0)|S_i(1)=s\}, E\{T_i(1)|S_i(1)=s\}) dF_S(s) / \Pr[S_i(1) > S_i(0)]$$

and

$$EDE \equiv E[h(E\{T_i(0)|S_i(1)\}, E\{T_i(1)|S_i(1)\})|S_i(1) = S_i(0)] \\ = h(E\{T_i(0)|S_i(1)=0\}, E\{T_i(1)|S_i(1)=0\})$$

are the expected associative and dissociative effects, and F_S is the CDF of $S_i(1)$. Values of $PAE \leq 0.5$ suggest a biomarker has no surrogate value, while biomarkers with some surrogate value will have $PAE \in (0.5, 1]$. The convention $|0|/(|0|+|0|) = 0.5$ is used to allow for the special case where $EAE = 0$ and $EDE = 0$. Note $PAE = 1$ if and only if $EDE = 0$ and $EAE \neq 0$, which occurs when (2) holds and $\Pr[S_i(1) > C] > 0$. In other words, PAE attains the upper bound of 1 when $S_i(1)$ is an SoP.

Following Follmann (2006), we model the transmission probability by

$$p(Z, S_i(1), W_i; \beta) = \Phi\{\beta_1 + \beta_2 Z + \beta_3 S_i(1) + \beta_4 Z S_i(1) + \beta_5 W_i\}, \quad (3)$$

where Φ is the standard normal CDF, W_i is some baseline covariate that is correlated with $S_i(1)$, and $\beta \equiv (\beta_1, \dots, \beta_5)$. Under this model, S will be an SoP if $\beta_2 = 0$ and $\beta_4 \neq 0$. Setting $h(x, y) = \Phi^{-1}(1/y) - \Phi^{-1}(1/x)$ yields

$$PAE = |\beta_2 + \kappa\beta_4| / \{|\beta_2| + |\beta_2 + \kappa\beta_4|\}, \quad (4)$$

where $\kappa \equiv E\{S_i(1)|S_i(1) > 0\} / \Pr[S_i(1) > 0]$. Letting F_W denote the CDF of W_i , the two curves $p(z, s; \beta) \equiv \int p(z, s, w; \beta) dF_W(w)$ for $z = 0, 1$ also inform about the surrogate value of S , with larger $|\beta_4|$ and smaller $|\beta_2|$ reflecting greater surrogate value.

For subject i , let $T_i \equiv \min\{T_i(0)(1-Z_i) + T_i(1)Z_i, C_i\}$ denote the observed number of exposures during the experiment where C_i denotes the right censoring time, i.e., the maximum allowable number of exposures. Let δ_i equal 1 (0) if subject i is infected (uninfected) by the end of the study and let $S_i \equiv S_i(0)(1-Z_i) + S_i(1)Z_i$ denote the observed immune response. Suppose we observe n iid copies of $O_i \equiv (Z_i, T_i, \delta_i, S_i, W_i)$. Letting $G(s|W)$ be the conditional distribution of $S_i(1)$ given W_i , the conditional likelihood is $L(\beta, G) \equiv \prod_{i=1}^n f(O_i; \beta, G)$ where

$$f(O; \beta, G) \equiv \phi(1, S, W, T, \delta; \beta) \left\{ \int \phi(0, s, W, T, \delta; \beta) dG(s|W) \right\}^{1-Z},$$

and $\phi(Z, S, W, T, \delta; \beta) \equiv \{1 - p(Z, S, W; \beta)\}^{T-\delta} p(Z, S, W; \beta)^\delta$.

Maximum “estimated likelihood” (Pepe and Fleming 1991) or “pseudolikelihood” (Liang and Self 1996) can be used for inference regarding β and PAE . As in Follmann (2006) and Gilbert and Hudgens (2008), we assume $(S(1), W)$ arise from a bivariate normal distribution with means (μ_S, μ_W) , variances (σ_S^2, σ_W^2) and correlation ρ , with left censoring of values of $S(1)$ or W below 0. We estimate $v \equiv (\mu_S, \mu_W, \sigma_S^2, \sigma_W^2, \rho)$ and hence G using $\{(S_i, W_i): Z_i = 1\}$ and $\{W_i: Z_i = 0\}$. Conditional on $\hat{G}$, the maximum estimated likelihood estimator (MELE) $\hat{\beta}$ is obtained by maximizing $\log\{L(\hat{G}, \hat{\beta})\}$ with respect to β . The estimator $\widehat{PAE}$ is computed by evaluating (4) at $\hat{v}, \hat{\beta}$. The bootstrap is used to estimate the standard errors of the estimators. A parametric bootstrap test (Davison and Hinkley 1997) is employed to assess $H_0: PAE = 0.5$ versus $H_A: PAE > 0.5$, i.e., to test the null S has no surrogate value versus the alternative S has some surrogate value. To conduct the parametric bootstrap test (PBT), B samples of size n are randomly sampled from the fitted null model using $\hat{v}$ and $\hat{\beta}_0$, the MELE of β under H_0 . For each sample, PAE is estimated and a one-sided test of size α is then conducted by comparing $\widehat{PAE}$ with the $(1 - \alpha) \times 100$ th percentile of the B bootstrap estimates of PAE .

3.2 Simulation study

A simulation study was conducted to assess whether sample sizes typical of RLD challenge studies provide adequate power to detect immune responses with high surrogate value. Data were generated assuming: m NHPs per arm; a maximum number of exposures per NHP of $C_i = 30$ for all i ; an average probability of infection per exposure for controls of $p(0) = 0.5$; a leaky vaccine effect with $VE_S = 0.8$; and model (3). Values of $(S_i(1), W_i)$ were generated by randomly sampling from a bivariate normal with $\mu_S = \mu_W = 0.5$, $\sigma_S^2 = \sigma_W^2 = 0.15$, and $\rho = 0.5, 0.7$ or 0.9 . Values of $S_i(1)$ or W_i below 0 were set equal to 0. Under this parameterization, approximately 10% of vaccinated animals have no (or undetectable) immune response. To reflect immune responses with varying surrogate values, simulated data sets were generated under three scenarios. For the first scenario, $\beta = (.38, -.33, -.63, -4.57, -.1)$ such that $PAE =$

0.9. Values for β for the remaining two scenarios were chosen such that $PAE = 0.5$ or 0.7 , with $p(1, S_i(1), W_i; \beta)$ the same as in the first scenario. For each scenario, 500 simulated data sets were generated. For each simulated data set, the MELEs $\hat{\beta}$ and $\widehat{PAE}$ were computed, their standard errors were estimated, and the PBT of $H_0: PAE = 0.5$ was conducted at the $\alpha = 0.1$ significance level.

Simulation results are given in Table 3 and Figure 3. The MELE $\widehat{PAE}$ is positively biased when $PAE = 0.5$ or 0.7 and negatively biased when $PAE = 0.9$, although the magnitude of the bias is negligible for $m > 10$. Likewise, the estimated transmission probability curves $p(0, s; \hat{\beta})$ and $p(1, s; \hat{\beta})$ exhibit minimal bias. The PBT has approximately the nominal size overall and adequate power for the high surrogate value scenario (i.e., $PAE = 0.9$) when there are 20 NHPs per arm. For each combination of PAE , ρ , and m in Table 3, we also estimated the power to detect a CoR, i.e., an association between $S_i(1)$ and $T_i(1)$, based on fitting a probit model to simulated data from the vaccine arm only. For all scenarios considered, the power to detect a CoR was at least 0.97, demonstrating that larger studies are needed to detect an immune marker with high surrogate value than to detect a CoR.

4 Discussion

Despite limited sample size, RLD challenge studies can inform about a vaccine's mechanism of protection and magnitude of effect. In particular, using discrete time survival models, we show that maximum likelihood methods can afford accurate and precise estimates of vaccine efficacy. While determining a vaccine's mechanism of protection is difficult in human studies, our results demonstrate that careful experimental design of challenge studies can lead to correct determination of the type of mechanism. We also consider a generalization of these models that allows for heterogeneous transmission probabilities. Similar extensions could easily be made to incorporate baseline covariates or allow for the possibility that a subset of NHPs are naturally immune to infection.

Our results also indicate it is possible to reliably evaluate potential immune SoPs in this setting. Properly designed RLD challenge studies can be adequately power to detect CoRs, and, to a lesser extent, immunological biomarkers with high surrogate value. These studies can also yield accurate estimates of surrogate value, based on the estimated transmission probability curves or functionals thereof such as PAE . However, results from our simulation study should be interpreted with caution for several reasons. First, 20 NHPs per arm are needed to have sufficient power to detect an immune marker with high surrogate value. Although such sample sizes are not the norm in this setting, several RLD challenge studies at least this large are being planned or conducted presently (John Mascola, personal communication). Second, the model assumed by the SoP analysis was correct, which in practice will rarely if ever be the case (discussed further below). Third, the assumed correlation of at least 0.5 between the HIV vaccine immune response $S_i(1)$ and the baseline predictor W_i may or may not be realistic. There are human studies of vaccines suggesting correlations of this magnitude may be realistic. For example, a study that vaccinated 75 individuals simultaneously with hepatitis A and B vaccines showed a linear correlation of 0.85 among A-specific and B-specific antibody titers (Czeschinski, Binding, and Witting, 2000), demonstrating that $W_i =$ hepatitis A titer may be an excellent baseline predictor for $S_i(1) =$ hepatitis B titer, and vice versa. Further data are needed from NHPs on realistic assumptions about the joint distribution of $S_i(1)$ and W_i to calibrate additional simulation studies of power to detect immune responses with high surrogate value.

The proposed approach to evaluating possible SoPs entails assuming SUTVA, ignorable treatment assignment, and a parametric regression model for the transmission probabilities. As discussed in Section 3.1, these first two assumptions should hold in RLD challenge studies.

The appropriateness of the probit model (3) will be more difficult to assess without additional information on $S_i(1)$. Some elaborations in the design of RLD challenge studies might be helpful in this regard. For example, in addition to the baseline predictor design studied in Section 3, Follman (2006) also proposed a second design, “closeout placebo vaccination” or CPV. Using this approach, controls who remain uninfected by the end of study would receive the HIV vaccine and their subsequent immune response would be measured. In the RLD challenge study setting, CPV may not be feasible since most, if not all, control animals are often infected after repeated challenges, e.g., see Subbarao et al. (2006) and Ellenberger et al. (2006). If this scenario is anticipated, a possible variation on the CPV design would be to vaccinate those NHPs randomized to the control arm that remain uninfected after a specific number of exposures. Further research is needed on applying CPV and variations therein to the RLD challenge setting.

We caution that the scope of inference drawn regarding the surrogate value of candidate specific SoPs should be limited to settings similar to the study at hand. A single RLD challenge study typically will not provide sufficient information for extrapolation to different vaccine formulations or human populations. Such inferences generally require conduct of additional studies. We refer the reader to Gilbert and Hudgens (2008) and Qin et al. (2007) for further related discussion of SoP assessment in vaccine studies.

In contrast to SoPs, smaller RLD challenge studies can provide adequate power to detect a CoR. For example, in our simulation study with 10 NHPs per arm, there was at most 50% power to detect a SoP with high surrogate value, whereas the power to detect a CoR was greater than 95%. In addition to requiring fewer NHPs, evaluation of potential CoRs does not require obtaining a baseline covariate W correlated with $S(1)$. These results are in concert with the well-known principle in the surrogate endpoint literature that establishing a valid surrogate requires substantially more evidence than merely determining a correlate.

Acknowledgments

This work was supported by NIH grant R01 AI054165-01. The authors thank Chih-Da Wu for his helpful comments and fitting the heterogeneous transmission probability model.

References

- Chick SE, Barth-Jones DC, Koopman JS. Bias reduction for risk ratio and vaccine effect estimators. *Statistics in Medicine* 2001;20:1609–1624. [PubMed: 11391691]
- Czeschinski P, Binding N, Witting U. Hepatitis A and hepatitis B vaccinations: immunogenicity of combined vaccine and of simultaneously or separately applied single vaccines. *Vaccine* 2000;18:1074–1080. [PubMed: 10590328]
- Davison, AC.; Hinkley, DV. *Bootstrap Methods and Their Application*. Cambridge University Press; 1997.
- Ellenberger D, Otten RA, Li B, Rodriguez V, Sariol CA, Martinez M, Monsour M, Wyatt L, Hudgens MG, Kraiselburd E, Moss B, Robinson H, Folks T, Butera S. HIV-1 DNA/MVA vaccination reduces the per exposure probability of infection during repeated mucosal SHIV challenges. *Virlogy* 2006;352:216–225.
- Farrington, CP. Communicable diseases. In: Armitage, P.; Colton, T., editors. *Encyclopedia of Biostatistics*. New York: Wiley; 1998. p. 795-815.
- Follmann D. Augmented designs to assess immune response in vaccine trials. *Biometrics* 2006;62:1161–1169. [PubMed: 17156291]
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002;58:21–29. [PubMed: 11890317]
- Gilbert PB. Interpretability and robustness of sieve analysis models for assessing HIV strain variations in vaccine efficacy. *Statistics in Medicine* 2001;20(2):263–279. [PubMed: 11169601]

- Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics*. 2008 In press.
- Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, Sinangil F, Burke D, Berman PW. Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases* 2005;191:666–77. [PubMed: 15688279]
- Halloran ME, Haber M, Longini IM. Interpretation and estimation of vaccine efficacy under heterogeneity. *American Journal of Epidemiology* 1992;136:328–343. [PubMed: 1415152]
- Halloran ME, Longini IM, Struchiner CJ. Design and interpretation of vaccine field studies. *Epidemiological Reviews* 1999;21:73–88.
- Jewell NP. On the bias of commonly used measures of association for 2×2 tables (C/R: V45 p1030–1032). *Biometrics* 1986;42:351–358.
- Liang KY, Self SG. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society, Series B: Methodological* 1996;58:785–796.
- Longini IM, Halloran ME. A frailty mixture model for estimating vaccine efficacy. *Applied Statistics* 1996;45:165–173.
- Pepe MS, Fleming TR. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* 1991;86:108–113.
- Qin L, Gilbert P, Corey L, McElrath M, Self S. A framework for assessing immunological correlates of protection in vaccine trials. *Journal of Infectious Diseases* 2007;196:1304–1312. [PubMed: 17922394]
- Regoes RR, Longini IM, Feinberg MB, Staprans SI. Preclinical assessment of HIV vaccines and microbicides by repeated low-dose virus challenges. *PLoS Medicine* 2005;2(8):e249. [PubMed: 16018721]
- Self SG, Liang K. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 1987;82:605–610.
- Smith PG, Rodrigues LC, Fine PEM. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *International Journal of Epidemiology* 1984;13:87–93. [PubMed: 6698708]
- Subbarao S, Otten R, Ramos A, Jackson E, Monsour M, Bashirian S, Kim C, Johnson J, Soriano V, Hudgens MG, Butera S, Janssen R, Paxton L, Greenberg A, Folks T. Chemoprophylaxis with Tenofovir Disoproxil Fumarate provided partial protection against Simian Human Immunodeficiency Virus infection in macaques given multiple virus challenges. *Journal of Infectious Diseases* 2006;194:904–11. [PubMed: 16960777]
- UNAIDS. Geneva: Dec. 2007 AIDS epidemic update. www.unaids.org
- Weinberg CR, Gladen BC. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* 1986;42:547–560. [PubMed: 3567288]

Figure 1.

Left panel: Nonparametric (solid line) and parametric (dotted line) estimates of the survival functions based on data from Ellenberger et al. (2006) and fitted leaky vaccine model from Table 1. Right panel: Nonparametric estimates of the complementary log-log survival functions.

Figure 2. Difference in nonparametric estimates of the complementary log-log survival functions between the vaccine and control arms for the observed data (bold lines) and 25 simulated data sets (gray lines) for each of the four different mechanism of protection models.

Figure 3.

Results from simulation study described in Section 3.2 with $m = 20$ NHPs per arm. Solid lines denote the true transmission probability curves $p(z, s; \beta)$ for $z = 0$ (upper line) and $z = 1$ (lower line). Dotted lines depict the mean of the estimated curves $p(z, s; \hat{\beta})$ over 500 simulated data sets.

Table 1

Maximum likelihood results using data from Ellenberger et al. (2006).

Model	θ	$\hat{\phi}$	$\hat{p}$	$\hat{V}_{E,S}$	Log-likelihood	AIC
Null	-	-	0.11	-	-79.07	160.14
All-or-none	0.22	-	0.16	0.22	-76.03	156.05
Leaky	-	0.36	0.20	0.64	-75.38	154.76
Mixture	0.16	0.53	0.20	0.56	-75.04	156.08

Table 2

Simulated power $\times 100\%$ to reject $H_0: VE_S = 0$. Each table entry is based on applying the likelihood ratio test (LRT) to 10,000 simulated data sets generated assuming $p = 0.2$ transmission probability in the control arm, sample size m per arm, and maximum of 20 challenges per NHP. The LRTs were conducted assuming the correct vaccine mechanism of protection at the $\beta = 0.1$ significance level.

		VE_S									
Mechanism	m	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Leaky	5	11	14	20	26	35	47	60	74	89	98
	10	11	15	24	35	49	64	80	93	99	100
	15	10	17	27	41	59	77	90	98	100	100
	20	10	19	31	48	67	85	96	99	100	100
All-or-none	5	4	34	57	73	85	92	97	98	100	100
	10	7	53	79	92	97	99	100	100	100	100
	15	8	64	89	97	100	100	100	100	100	100
	20	8	71	95	99	100	100	100	100	100	100

Table 3

Results from simulation study described in Section 3.2. Each table entry is based on 500 simulated data sets with m NHPs per arm. ρ is the linear correlation of the simulated bivariate normal variables giving rise to W and $S(1)$. Bias is the median bias and SE is empirical standard error of $\widehat{PAE}$. SEE is the median of the bootstrap standard error estimates. EP is the empirical power for the parametric bootstrap test of $H_0:PAE = 0.5$ versus $H_A:PAE > 0.5$ at level $\alpha = 0.1$.

		$m = 10$					$m = 15$					$m = 20$					
PAE	ρ	Bias	SE	SEE	EP	Bias	SE	SEE	EP	Bias	SE	SEE	EP	Bias	SE	SEE	EP
0.50	0.5	0.10	0.22	0.20	0.14	0.05	0.19	0.18	0.09	0.04	0.17	0.16	0.09	0.04	0.17	0.16	0.09
	0.7	0.08	0.21	0.19	0.13	0.03	0.19	0.17	0.07	0.03	0.16	0.16	0.10	0.03	0.16	0.16	0.10
	0.9	0.06	0.20	0.18	0.11	0.06	0.17	0.16	0.11	0.02	0.15	0.15	0.09	0.02	0.15	0.15	0.09
0.70	0.5	0.05	0.17	0.16	0.26	0.06	0.14	0.14	0.36	0.05	0.14	0.13	0.50	0.05	0.14	0.13	0.50
	0.7	0.03	0.16	0.15	0.29	0.05	0.15	0.14	0.35	0.06	0.14	0.13	0.53	0.06	0.14	0.13	0.53
	0.9	0.04	0.16	0.15	0.32	0.04	0.14	0.14	0.47	0.05	0.13	0.12	0.62	0.05	0.13	0.12	0.62
0.90	0.5	-0.09	0.14	0.14	0.36	-0.05	0.11	0.12	0.55	-0.03	0.11	0.11	0.80	-0.03	0.11	0.11	0.80
	0.7	-0.09	0.15	0.14	0.40	-0.04	0.12	0.12	0.65	-0.03	0.10	0.11	0.83	-0.03	0.10	0.11	0.83
	0.9	-0.08	0.15	0.14	0.50	-0.05	0.11	0.12	0.77	-0.03	0.10	0.10	0.91	-0.03	0.10	0.10	0.91