Biometrics 62, 1161–1169
December 2006

DOI: 10.1111/j.1541-0420.2006.00569.x

# Augmented Designs to Assess Immune Response in Vaccine Trials

**Dean Follmann**

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases,
6700B Rockledge Drive MSC 7609, Bethesda, Maryland 20892, U.S.A.
*email:* dfollmann@niaid.nih.gov

Summary. This article introduces methods for use in vaccine clinical trials to help determine whether the immune response to a vaccine is actually causing a reduction in the infection rate. This is not easy because immune response to the (say HIV) vaccine is only observed in the HIV vaccine arm. If we knew what the HIV-specific immune response in placebo recipients would have been, had they been vaccinated, this immune response could be treated essentially like a baseline covariate and an interaction with treatment could be evaluated. Relatedly, the rate of infection by this baseline covariate could be compared between the two groups and a causative role of immune response would be supported if infection risk decreased with increasing HIV immune response only in the vaccine group. We introduce two methods for inferring this HIV-specific immune response. The first involves vaccinating everyone before baseline with an irrelevant vaccine, for example, rabies. Randomization ensures that the relationship between the immune responses to the rabies and HIV vaccines observed in the vaccine group is the same as what would have been seen in the placebo group. We infer a placebo volunteer's response to the HIV vaccine using their rabies response and a prediction model from the vaccine group. The second method entails vaccinating all uninfected placebo patients at the closeout of the trial with the HIV vaccine and recording immune response. We pretend this immune response at closeout is what they would have had at baseline. We can then infer what the distribution of immune response among placebo infecteds would have been. Such designs may help elucidate the role of immune response in preventing infections. More pointedly, they could be helpful in the decision to improve or abandon an HIV vaccine with mediocre performance in a phase III trial.

Key words: AIDS; Causal inference; Correlate of protection; Counterfactual; HIV; Missing data; Principal stratification; Surrogate endpoint.

## 1. Introduction

A vaccine contains innocuous material that provokes a response by the adaptive immune system. Following vaccination, the immune system mounts a multifaceted, and exquisitely specific, counterattack based on two types of white blood cells, B-lymphocytes and T-lymphocytes. These cells respond to specific proteins of the vaccine material, proliferate, and wait to subsequently attack either floating microbes or infected cells that display such peptides. B-lymphocytes produce antibodies that recognize proteins in the outer surface of the virus and neutralize their ability to infect cells. T-lymphocytes produce cells that either kill or aid in killing infected cells. The magnitude of each component of the adaptive immune response to the vaccine can be measured. Vaccine development focuses on inducing a strong, measurable immune response while ensuring that the vaccine is safe (see, e.g., Halloran, 1998; Nabel, 2001; or Chan, Wang, and Heyse, 2003).

Establishing the role of vaccine-induced immune response on actual protection of infection and disease is an important open problem in vaccine studies (Halloran, 1998). A "correlate of protection" is the threshold for immune response, say $x_p$, beyond which infections and disease do not occur (Lachenbruch et al., 2000). Methods for estimating such a threshold are discussed in Carey, Barker, and Platt (2001), Chan et al. (2002), and Plikaytis and Carlone (2005). However, when immune response only occurs in the vaccinated group, validation of a correlate of protection, or more generally validation of immune response as a true surrogate with a causative role, is problematic (Chan et al., 2003). The use of Prentice's criteria to establish surrogacy, conditional independence of treatment, and outcome given the surrogate (Prentice, 1989) breaks down here because immune response to the vaccine basically only occurs in the vaccine group and thus the value of the surrogate basically identifies the treatment group. Strictly speaking, one cannot know whether the measured immune responses, or other unmeasured vaccine-induced changes, are actually responsible for an efficacious vaccine. For example, it could be that those individuals who achieve $x_p$ in response to a weak vaccine are more intrinsically fit than others so that even if a more powerful vaccine achieved $x_p$ in everyone, not all would be protected.

That this might be an actual problem was demonstrated in VAX004, the first phase III trial of an HIV vaccine (Gilbert et al., 2005; The rgp120 HIV Vaccine Study Group, 2005). Overall, the vaccine was not effective, with infection rates of 0.067 and 0.070, respectively, in the vaccine and placebo groups based on 5403 volunteers. However, the antibody re-

**Table 1**

*The relative hazard of infection, based on a Cox model, as a function of antibody response to the HIV vaccine, which is only measured in the vaccine group. It seems the vaccine-induced antibodies are doing their job.*

| | Quartile of antibody response following HIV vaccination | | | |
|---|---|---|---|---|
| Group | Weak | Modest | Good | Best |
| Vaccine | 1.00 | 0.35* | 0.28* | 0.22* |

*$p < 0.05$.

sponse to the HIV vaccine was strongly associated with infection risk in the vaccine group. Tables 1 and 2 provide the relative hazard of infection as a function of antibody response quartiles, first within the vaccine group and then when the placebo group is used as a control (see Gilbert et al., 2005). Because antibody response to the HIV vaccine is only measured in the vaccine group, Table 2 has question marks in the placebo cells—we do not know what HIV immune response they would have had, had they been vaccinated.

Two hypotheses were postulated to explain these results (Gilbert et al., 2005; Graham and Mascola, 2005). The first was that antibody response is identifying volunteers with different constitutional ability to avoid infection but the vaccine-induced immune response had no causative role. We call this the association hypothesis. The second was that the vaccine caused infections in those with the weakest immune response and prevented infections in those with the strongest immune response. We call this the causation hypothesis. As it stands, neither of these hypotheses can be evaluated on the basis of data.

In this article we introduce two new designs to help understand the role of immune response in vaccines. These designs can discriminate between the two hypotheses outlined above. The first design is to inoculate everyone in both arms prior to randomization with an irrelevant vaccine, say rabies. We call this *baseline irrelevant vaccination* (BIV), and let $W_0$ be the immune response to the rabies vaccine at baseline. Also, we define $X_0$ as the immune response to the HIV vaccine, which is measured just after randomization in the vaccine group. Randomization ensures that the relationship between $W_0$, $X_0$ observed in the vaccine group is the same in the placebo group. Based on this relationship, the observed $W_0$ of a placebo participant can be used to infer his $X_0$. Figure 1 illustrates how $W_0$ can be used to impute $X_0$ in the placebo group when they are very highly correlated ($\rho = 0.98$). It is important to note that a rabies vaccine is not required—any baseline measurement that correlated well with $X_0$ would work, but an irrelevant vaccination is a good choice. This type of thinking to predict a post-randomization characteristic only observed in the treatment group has been used before in heart disease (see, e.g., Follmann, 2000, or Hallstrom et al., 2001).

**Table 2**

*When we calculate the relative hazard for the four quartiles compared to the placebo group, a different picture emerges* (*Gilbert et al.,* 2005). *The numbers provide the hazard relative to the overall placebo group, while the ?'s emphasize that immune response following HIV vaccination is not measurable in the placebo group and thus the relative hazards are unknown.*

| | Quartile of antibody response following HIV vaccination | | | | |
|---|---|---|---|---|---|
| Group | Weak | Modest | Good | Best | Overall |
| Placebo | ? | ? | ? | ? | 1.00 |
| Vaccine | 1.86* | 0.99 | 0.99 | 0.81 | |

*$p < 0.05$.

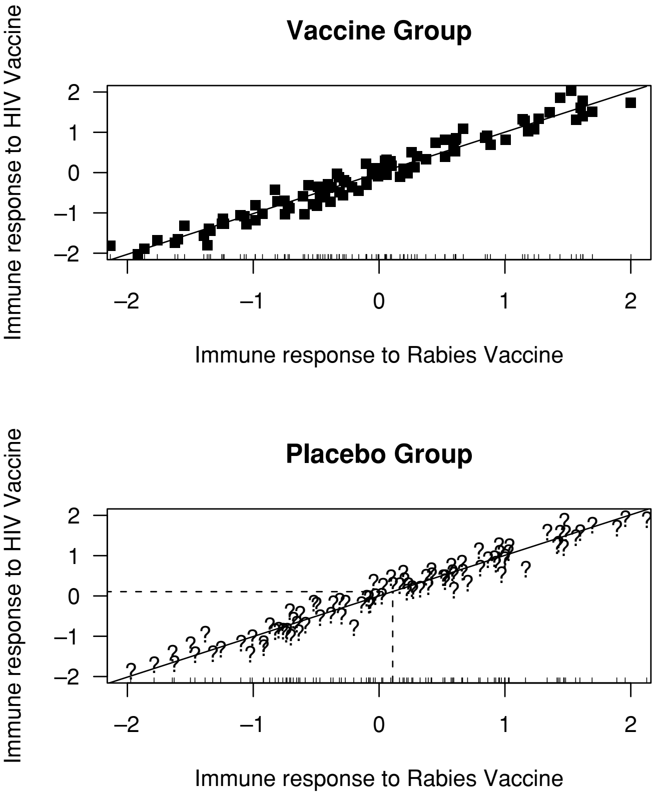


**Figure 1.** Made-up scatterplot illustrating imputation of the immune response to an HIV vaccine ($X_0$) in the placebo group based on the observed immune response to a rabies vaccine ($W_0$) for a single patient. The bivariate distribution between $X_0$, $W_0$ is observed in the vaccine group. Randomization assures that this distribution and regression line also apply to the placebo group. While $X_0$ cannot be observed in the placebo group, $W_0$ can and provides the basis for imputation. A very high correlation between $X_0$, $W_0$ is used to illustrate the concept.

The second way to get at $X_0$ in the placebo group would be to vaccinate all the uninfected placebo recipients at the closeout of the trial with the HIV vaccine and then measure their immune response, say $X_C$. If we make the assumption that $X_C$ is the same as $X_0$, we effectively obtain $X_0$ in many.

**Table 3**

*Hypothetical data set of a trial where* 800 *patients are randomized. The vaccine group has an immune response to the HIV vaccine that is measured just after randomization/vaccination. The placebo volunteers who remain uninfected are vaccinated at the end of the study and immune response is measured then. Bold numbers are directly observed, italicized numbers are inferred. Randomization assures that roughly* 100 *placebo patients would be in each quartile, as occurred in the vaccine group. In this example, consistent with the association hypothesis, the vaccine has no overall effect but identifies patients with an intrinsic ability to avoid infection.*

| | Quartile of antibody response following HIV vaccination | | | | |
|---|---|---|---|---|---|
| Group | 1st | 2nd | 3rd | 4th | Total |
| Vaccine | | | | | |
| Total | **100** | **100** | **100** | **100** | **400** |
| Infected | **30** | **15** | **10** | **5** | **60** |
| Uninfected | **70** | **85** | **90** | **95** | **340** |
| Placebo | | | | | |
| Total | ±*100* | ±*100* | ±*100* | ±*100* | **400** |
| Infected | ±*29* | ±*16* | ±*9* | ±*4* | **58** |
| Uninfected | **71** | **84** | **91** | **96** | **342** |

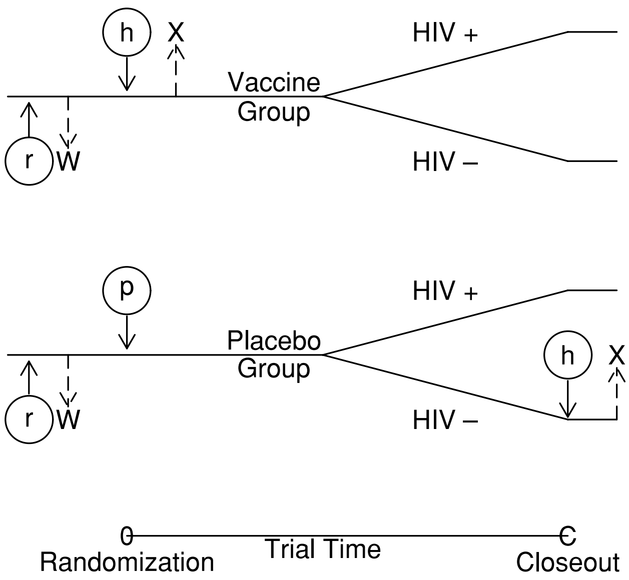


**Figure 2.** Schematic representation of augmented designs. Circles and lowercase letters denote inoculations, immuneresponse is denoted by capital letters. Under a traditional design, patients are vaccinated either with HIV vaccine (h) or placebo (p) and immune response to the HIV vaccine ($X = X_0$) is measured shortly thereafter in the vaccine group. Under BIV, both groups are vaccinated against rabies (r) and the immune response to rabies vaccine ($W = W_0$) is measured prior to randomization. Under CPV, placebo patients who are uninfected at the end of the trial receive HIV vaccine at closeout and their immune response is measured then ($X = X_C$).

We call this *closeout placebo vaccination* (CPV). Table 3 provides hypothetical data illustrating how CPV can be used to suggest that $X_0$ is associated with constitutional ability to remain uninfected, but has no causative role.

Figure 1 and Table 3 are meant to informally illustrate how to infer $X_0$ in the placebo group. In the sequel, we develop formal methods that rely on the thinking of counterfactuals, causal inference, and principal stratification. We also describe some simple methods, investigate performance of different methods by simulation, and discuss some more elaborate approaches.

## 2. Model-Based Approach

Suppose that $n$ patients per group are randomized to placebo or vaccine. Prior to randomization, all patients receive a rabies vaccine and the immune response to rabies vaccine ($W_0$) is measured before randomization. Patients are then randomized to either a placebo or HIV vaccine injection and shortly thereafter, immune response to the HIV vaccine ($X_0$) is measured in the vaccine group. At the closeout or end of the trial, all uninfected placebo recipients receive the HIV vaccine and shortly thereafter, immune response to this vaccine is measured ($X_C$). Let $Y$ be the infection indicator and $Z$ be the vaccine indicator. A schematic representation of a vaccine trial augmented with BIV and CPV is given in Figure 2.

Our approach to using these data is perhaps best described using counterfactual reasoning (Rubin, 1974, 1977, 1978; Halloran and Struchiner, 1995) and principal stratification (Frangakis and Rubin, 2002). First, let $W_{0i}$ be the baseline rabies-specific adaptive immune response for patient $i$. This is seen in everyone. The response to HIV vaccination is different. One can write $X_{0i}(z)$ as the (post) baseline HIV-specific immune response to HIV vaccination. We call $X_{0i}(0)$, $X_{0i}(1)$ potential covariates; $X_{0i}(1)$ is measured in vaccine recipients while $X_{0i}(0)$ would be 0 in nearly everyone. We say that $X_{0i}(1)$ is *realized* in the vaccine group and *unrealized* in the placebo group. Using the terminology of Frangakis and Rubin (2002), $X_i(1) = x$, $X_i(0) = 0$ defines a principal stratum indexed by $x$. Principal strata are a classification of subjects defined by the potential values of a post-treatment variable under each of the treatments being considered. They also call $X_0(1)$ a principal surrogate and distinguish it from a "statistical" surrogate, which for our setup would be $X^{\text{obs}} = X_0(1)Z + X_0(0)\ (1 - Z)$. We next define $Y_i(z)$ as the outcome for person $i$ following treatment $z$. We call the pair $Y_i(0)$, $Y_i(1)$ potential outcomes. We also define $X_{Ci}(z, y)$ as the closeout HIV-specific adaptive immune response for person $i$ when given treatment $z$ and following outcome $y$. Only $X_{Ci}(0, 0)$ is measured and meaningful:

We make the following simplifying assumptions:

- All patients receive the assigned injections so there is no noncompliance.
- There are no missing data; $W_0$, $Y_0$ are measured on everyone, $X_0$ is measured on all vaccinees, and $X_C$ is measured on all placebo uninfecteds.
- No infections occur between the time of randomization and when $X_0$ is measured, say the interval [0, $m$].

The first two are for simplicity and can be relaxed. For example, if there is some noncompliance but it is governed by an independent random mechanism, our methods could be applied to just the compliers. With data missing completely at random the methods can be applied directly to the observed data. If the data are missing at random, methods that incorporate covariates associated with missingness can be used. The last assumption is more likely to be met if $m$ is small. If a few infections occur in [0, $m$], an analysis that throws them out may be acceptable. We discuss how to modify our approach to incorporate infections during [0, $m$] in Section 6.

We next specify probit models for the effect of the "baseline covariate" $X_0(1)$ on the probability of infection in both groups:

$$\begin{aligned} p_z(x) &= P\{Y_i(z) = 1 \mid Z_i = z, X_{0i}(1) = x\} \\ &= \Phi(\beta_0 + \beta_1 z + \beta_2 x + \beta_3 zx), \end{aligned} \tag{1}$$

where $\Phi(\,)$ is the standard normal c.d.f. (cumulative distribution function). This equation specifies a model for a standard covariate by treatment interaction for a clinical trial. The probit is handy because it is easy to integrate over $x$, which we will need to do later. Note that (1) assumes that $W_0$ has no effect on $Y(z)$ once $X_0(1)$ and $Z$ are in the model. This can also be relaxed, as we discuss in Section 5.

Different causal estimands can be used to quantify the effect of the vaccine as a function of $X_0(1)$. For example, following Hudgens and Halloran (2004) we define vaccine efficacy as

$$\begin{aligned} 1 - \frac{E\{Y_i(1) \mid X_{0i}(1) = x\}}{E\{Y_i(0) \mid X_{0i}(1) = x\}} &= VE(x) \\ &= 1 - \frac{\Phi\{\beta_0 + \beta_1 + (\beta_2 + \beta_3)x\}}{\Phi(\beta_0 + \beta_2 x)}. \end{aligned}$$

With our probit model, a natural estimand is

$$\begin{aligned} &\Phi^{-1}[E\{Y_i(1) \mid X_{0i}(1) = x\}] - \Phi^{-1}[E\{Y_i(0) \mid X_{0i}(1) = x\}] \\ &\quad = \Delta_P(x) = \beta_1 + \beta_3 x. \end{aligned}$$

Note that when $\beta_3 = 0$, $\Delta_P(x)$ is free of $x$, this is not true for $VE(x)$.

If $X_{0i}(1)$ were observed in everyone, estimation would be straightforward. As $X_{0i}(1)$ is not observed in the placebo group, we require at least one of the following two assumptions to proceed:

- $X_{0i}(1)$ can be viewed as a baseline covariate or
$$\{X_{0i}(1) \mid W_{0i}, Z = 0\} \overset{D}{=} \{X_{0i}(1) \mid W_{0i}, Z = 1\}.$$
- For placebo uninfecteds, $X_{0i}(1) = x_i + U_1$ and $X_{Ci}(0, 0) = x_i + U_2$ where $U_1$ and $U_2$ are i.i.d. (independent and identically distributed) mean 0. We call this time constancy of immune response.

The first assumption is true by design in randomized trials and allows us to impute $X_{0i}(1)$ based on $W_{0i}$ in the placebo group. While technically measured post-randomization, this "post-baseline" covariate can be used as a baseline covariate. The second assumption allows us to replace $X_{0i}(1)$ with $X_{Ci}(0, 0)$ as a covariate in the probit model for placebo uninfecteds. Under the model $X = x + U$, one can think of $x$ as the true time constant immune response, which is observed subject to measurement error and our interest focuses on the regression of $Y$ on $X$. This assumption cannot be accepted uncritically as immune response can diminish with age, such as for herpes zoster, if the trial is long enough. Additionally, volunteers might get subinfectious exposures to a virus that modifies immune response. This is thought possible for HIV where commercial sex workers showed immune responses to HIV but remained uninfected. However even here, the assumption might hold if the immune response is effectively primed by subinfectious exposure pre-baseline and this response is maintained during the course of the trial. Additionally, this assumption can be examined, as we will discuss in Section 5.

Our final assumption allows us to easily integrate over the distribution of $X_0(1) \mid W_0$:

- The distribution of $X_0(1)$, $W_0$ is bivariate normal with moments $\mu_x$, $\mu_w$, $\sigma_x^2$, $\sigma_w^2$, $\rho$.

This assumption can also be relaxed but the integration would be more complicated.

To estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$, we use maximum likelihood. We begin by constructing a likelihood incorporating both BIV and CPV. The likelihood contribution for vaccinees is simple,

$$\prod_{i \in \mathcal{V}} p_1(x_{0i})^{y_i} \{1 - p_1(x_{0i})\}^{1-y_i},$$

where $\mathcal{V}$ is the set of vaccinees. For uninfected placebo volunteers we use $X_{Ci}$ in lieu of $X_{0i}$ and their contribution is

$$\prod_{i \in \mathcal{P}(U)} \{1 - p_0(x_{Ci})\}^{1-y_i},$$

where $\mathcal{P}(\mathcal{U})$ is the set of uninfected placebo recipients. In the placebo infecteds, $X_0(1)$ is missing and we need to integrate $p_0(X_0(1))$ over the distribution of $X_0(1) \mid W_0$ to obtain their likelihood contribution. Under our last assumption, it follows that $X_0(1) \mid W_0 = w$ is normal with mean $\mu_*(w_0) = \mu_x + \rho\sigma_x/\sigma_w(w_0 - \mu_w)$ and variance $\sigma_*^2 = \sigma_x^2(1 - \rho^2)$. The (integrated) probability of infection for a person with $W_0 = w_0$ is thus

$$p_0^*(w_0) = E[\Phi\{\beta_0 + \beta_2 X_0(1)\}] = \Phi\left\{\frac{\beta_0 + \beta_2\mu_*(w_0)}{\sqrt{1 + (\beta_2\sigma_*)^2}}\right\}.$$

The right-hand side obtains the result that $E[\Phi(a + U)] = \Phi\{(a + \mu)/\sqrt{1 + \sigma^2}\}$ for $U$ normal($\mu$, $\sigma^2$). The overall likelihood is thus

$$\begin{aligned} L_{BC}(\boldsymbol{\beta}) = &\left[\prod_{i \in \mathcal{V}} p_1(x_{0i})^{y_i}\{1 - p_1(x_{0i})\}^{1-y_i}\right] \\ &\times \left[\prod_{i \in \mathcal{P}(U)} \{1 - p_0(x_{Ci})\}\right] \left\{\prod_{i \in \mathcal{P}(I)} p_0^*(w_{0i})\right\}. \end{aligned}$$

Note that $p_0^*(w_{0i})$ depends on the moments of $X_0(1)$, $W_0$, which are unknown. We advocate estimating these moments using vaccine group data and regard them as fixed in $L_{BC}$. Because of this, the standard error estimates obtained by the Fisher information matrix are incorrect and we suggest using the nonparametric bootstrap method to obtain standard errors.

We can also construct likelihoods based on augmenting the usual design with BIV alone or CPV alone. These are, respectively,

$$L_B(\boldsymbol{\beta}) = \left[\prod_{i\in\mathcal{V}} p_0(x_{0i})^{y_i}\{1-p_0(x_{0i})\}^{1-y_i}\right] \times \left[\prod_{i\in\mathcal{P}} p_0^*(w_{0i})^{y_i}\left\{1-p_0^*(w_{0i})\right\}^{1-y_i}\right],$$

where $\mathcal{P}$ is the set of placebo recipients, and

$$L_C(\boldsymbol{\beta}) = \left[\prod_{i\in\mathcal{V}} p_0(x_{0i})^{y_i}\{1-p_0(x_{0i})\}^{1-y_i}\right] \times \left[\prod_{i\in\mathcal{P}(U)}\{1-p_0(x_{Ci})\}\right]\Phi\left\{\frac{\beta_0+\beta_2\mu_x}{\sqrt{1+(\beta_2\sigma_x)^2}}\right\}^{\#\mathcal{P}(I)},$$

where $\#\mathcal{P}(I)$ is the number of placebo infecteds. The last $\Phi(\,)$ in $L_C(\boldsymbol{\beta})$ is just the probability that a generic placebo patient is infected and equals $E\{\beta_0+\beta_1X_0(1)\}$, where $X_0(1)$ is normal $(\mu_x, \sigma_x^2)$. Based on the estimated $\beta$'s it is a simple matter to plug them into a causal estimand. Standard errors and confidence intervals for causal estimands can be computed from the bootstrap.

## 3. Closeout Placebo Vaccination Alone

The previous section outlined how BIV and CPV can be used to estimate the effect of immune response using a model and likelihood. In this section, we show how closeout placebo vaccination by itself can be used without a model to assess immune response. The approach is inspired by Tables 1 and 2 and Gilbert, Bosch, and Hudgens (2003).

Denote by $f_0(x)$ and $f_1(x)$ the densities of $X_0(1)$ for the placebo and vaccine groups, respectively. In each group we can decompose the distribution of immune response into a mixture of those who would/did become infected and those who would not/did not. Thus we can write the immune response densities in mixture form,

$$f_0(x) = f_0(x \mid Y=1)\theta_0 + f_0(x \mid Y=0)(1-\theta_0), \quad \text{and} \qquad (2)$$

$$f_1(x) = f_1(x \mid Y=1)\theta_1 + f_1(x \mid Y=0)(1-\theta_1), \qquad (3)$$

where $\theta_\ell$ is the true proportion of infected volunteers in group $\ell$. In the vaccine group the mixed density and the two constituent densities are directly estimable as is $\theta_1$. In the placebo group $\theta_0$ and $f_0(x \mid Y=0)$ are directly estimable, provided $\{X_0(1) \mid Y=0\} \overset{D}{=} (X_C \mid Y=0)$. To get $f_0(x \mid Y=1)$ we replace $f_0(x)$ with $f_1(x)$ and solve by subtraction.

With these arguments and Bayes' theorem, one can deduce that

$$p_0(Y=1 \mid x) = \theta_0\frac{f_0(x \mid Y=1)}{f_1(x)}, \quad \text{and} \qquad (4)$$

$$p_1(Y=1 \mid x) = \theta_1\frac{f_1(x \mid Y=1)}{f_1(x)}. \qquad (5)$$

The terms on the right-hand side can be estimated nonparametrically and thus so can the left-hand side.

Interestingly, the different conditional distributions of $X_0(1)$ can be compared to test the role of $X_0(1)$. To motivate these tests, consider Table 3. Suppose the counts in the placebo uninfected row were very similar over the four quartiles. This would suggest that *unrealized* potential immune response was unassociated with infection risk. Using the fact that $f_1(x) = f_0(x)$, the continuous analog to see whether the counts in the placebo uninfected row are similar can be written as

$$\mathrm{H}_0^2: f_0(x \mid Y=0) = f_1(x) \iff p_0(Y=1 \mid x) = \theta_0.$$

Note that if the probit model (1) is correct, then $\mathrm{H}_0^2$ is equivalent to $\beta_2 = 0$. Also note that $\mathrm{H}_0^2$ corresponds to the causation hypothesis that was suggested to explain Tables 1 and 2.

At the other extreme, suppose that the counts in the vaccine uninfected row were quite similar to the counts in the placebo uninfected row. This would suggest that immune response has no causative effect on infection. The continuous analog is

$$H_0^3: f_0(x \mid Y=0) = f_1(x \mid Y=0)$$
$$\iff p_0(Y=0 \mid x) \propto p_1(Y=0 \mid x).$$

Unlike $\mathrm{H}_0^2$, $\mathrm{H}_0^3$ does not correspond to $\beta_3 = 0$ even if (1) is correct, unless $\beta_1 = 0$. Note that $\mathrm{H}_0^3$ corresponds to the association hypothesis suggested to explain Tables 1 and 2.

Different methods could be used to test equality of the densities specified by $\mathrm{H}_0^2$ and $\mathrm{H}_0^3$ such as *t*-tests, rank tests, or Kolmogorov-type tests. For a *t*-test of $\mathrm{H}_0^2$, one compares all $X_{0i}(1)$'s in the vaccine group to the $X_{Ci}$'s of the placebo uninfecteds. For a *t*-test of $\mathrm{H}_0^3$, one compares the $X_{0i}(1)$'s of the vaccine uninfecteds to the $X_{Ci}$'s of the placebo uninfecteds.

## 4. Simulation

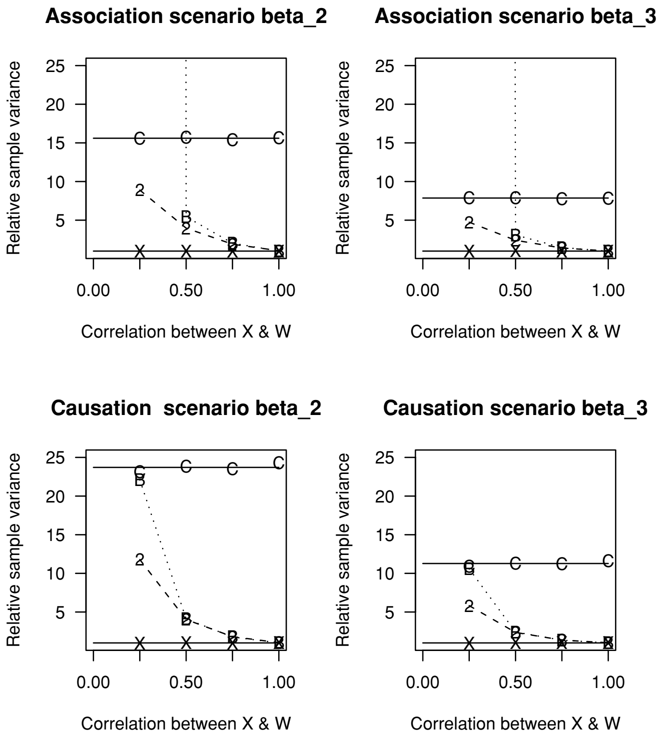
To assess these designs, we conducted a simulation under the model assumptions given in the previous section. We generated data where $P\{Y(z) = 1 \mid Z = z, X_0(1) = x\}$ is given by (1), and $W_0$, $X_0(1)$ are bivariate Gaussian with correlation $\rho$. We set $E[p_0\{X_0(1)\}] = \theta_0 = 0.10$ and $\theta_1 = 0.08$. We selected $\beta_2$, $\beta_3$ in terms of relative risk,

$$\frac{p_\ell\{Q(7/8)\}}{p_\ell\{Q(1/8)\}} = R_\ell,$$

where $Q(7/8)$, $Q(1/8)$ are the seventh and first octiles of the distribution of $X_0(1)$. Three scenarios were considered, chosen with the hazards of Tables 1 and 2 in mind:

- Association: Here $R_1 = R_0 = 0.2$, $\beta_3 = 0$, and $\Delta_P(x)$ is free of $x$.
- Causation: Here $R_0 = 1$, $R_1 = 0.2$, $\beta_2 = 0$, and $\Delta_P(x)$ depends on $x$.
- Both: Here $R_0 = 0.33$, $R_1 = 0.11$, $\beta_k < 0$, $k = 0, 1$, and $\Delta_P(x)$ depends on $x$.

For each simulated data set maximum likelihood using $L_{BC}$, $L_B$, and $L_C$ was used to estimate $\boldsymbol{\beta}$. We also constructed a probit likelihood based on observing $X_0(1)$ exactly in everyone. Estimates based on this likelihood correspond to an unattainable benchmark.


Association scenario beta_2

Association scenario beta_3

Causation scenario beta_2

Causation scenario beta_3

**Figure 3.** Sample variance of estimates of $\beta$ divided by the sample variance when the $X_0(1)$ is used. Estimates denoted by $B$, $C$, 2, and $X$ correspond to designs using BIV alone, CPV alone, BIV + CPV, and the impossible benchmark where $X_0(1)$ is known in everyone, respectively. For BIV alone when $\rho = 0.25$ the relative sample variance is enormous and off the chart for the Association scenario. One can extrapolate the behavior of the designs using CPV alone and the benchmark $\rho = 0$ as their behavior is free of $\rho$. Each symbol is based on 10,000 simulated trials.

The first set of simulations used 10,000 replications and varied by 0.25, 0.5, 0.75, 1. We do not evaluate $\rho = 0$ as the model using BIV alone is unidentifiable. Replications were not tallied when convergence was not attained, which was very rare except for BIV alone with $\rho = 0.25$ when the estimates did not converge 2–3% of the time.

Figure 3 provides the sample variance for the four estimates of $\beta$, divided by the sample variance when $X_0(1)$ is used, as a function of $\rho$ under the Association and Causation scenarios. Relative behavior of the different estimates is similar under the Both scenario and hence not reported. For the estimates using CPV (C) alone or the benchmark (X), the sampling variability is free of $\rho$. The sample variance with CPV alone is from nearly 10 times to almost 25 times larger than with the benchmark. The performance of BIV (B) alone depends profoundly on $\rho$ with $\rho = 0.25$ exhibiting extremely large sample variances for the Association scenario, and variances similar to CPV alone for the Causation scenario. For $\rho > 0.5$, BIV and CPV + BIV have similar variance ratios. We see that for large $\rho$ CPV is unnecessary and for small $\rho$ BIV performs poorly. As $\rho = 0.25$, both CPV and BIV are helpful.

Our second set of simulations evaluates power and is given in Table 4 with $n = 1000$ or 2500, $\rho = 0.25$ or 0.50, for the three scenarios Association, Causation, and Both. For the Wald tests, a nonparametric bootstrap standard error was calculated using the sample variance of 100 bootstrap resamples for each simulated trial. Resamples where convergence was not attained were thrown out, which was rare except for BIV alone with $\rho = 0.25$. As before, BIV alone with $\rho = 0.25$ had problems with convergence and these were exacerbated in the bootstrap resamples.

We begin by evaluating the Wald test. First, the benchmark has extremely high power, except for $\beta_3$ under scenario B with $n = 1000$. For CPV + BIV, power is generally good to excellent for all scenarios with $n = 2500$. For $n = 1000$, power is degraded, especially with $\rho = 0.25$. For BIV alone, power is similar to CPV + BIV for $\rho = 0.50$ and much worse for $\rho = 0.25$. Generally, power for CPV alone is much worse than for BIV alone with $\rho = 0.50$ and moderately better with $\rho = 0.25$. The power of the $t$-tests is usually similar to CPV alone and close to at least 0.50 for scenarios A and C with $n = 2500$.

We also did a few limited simulations to address specific issues. In practice, one might want to perform CPV on a fraction of the placebo uninfecteds. For scenario A, we compared the estimates using CPV alone, where $X_C$ was obtained in everyone to where it was obtained in 1/2, 1/4, or 1/10 of the placebo uninfecteds. The sampling variance for either $\hat{\beta}_2$ or $\hat{\beta}_3$ was about 60%, 300%, and 1000% larger than when $X_C$ was obtained in everyone, respectively. Second, we evaluated the procedures when the moments were set to their true values and not estimated. The sampling variance for CPV alone and for BIV alone was nearly halved when true values were used instead of estimated values. For CPV + BIV, the sampling variability was only modestly reduced. For larger trials, for example, $n = 8000$ with low event rates, the performance of CPV and BIV relative to BIV + CPV might be better than shown in Figure 3 and Table 4 as the estimated moments of $X_0(1)$, $W_0$ would be more reliable. It also suggests that one might want to consider use of a full likelihood. For example, for CPV alone uses

$$\begin{aligned} &L\big(\boldsymbol{\beta}, \mu_x, \sigma_x^2\big) \\ &\quad = L_C(\boldsymbol{\beta}) \prod_{i \in \mathcal{V}} \phi\big(x_{i0}; \mu_x, \sigma_x^2\big) \prod_{i \in \mathcal{P}(U)} f_0\big(x_{iC} \,\big|\, Y_i = 0; \mu_x, \sigma_x^2\big), \end{aligned}$$

where $\phi(x; \mu, \sigma^2)$ is the normal density and $f_0(x_{iC} \mid Y_i = 0; \mu_x, \sigma_x^2)$ is the density for uninfecteds, derived under (1) and a Gaussian model for $X_0(1)$.

In summary, the new designs can be efficient and powerful even with $n = 1000$ if $\rho > 0.5$. If $\rho$ is modest, a larger sample size is required to achieve strong power as CPV is necessary. If $\rho$ is large enough, CPV may be unnecessary, while if $\rho$ is too small, BIV alone may be useless. With $n = 2500$ we have excellent power for scenarios A and C with $\rho = 0.5$ using BIV alone and good to excellent power with the BIV + CPV combination with $\rho = 0.25$. Even with CPV alone, power is greater than 50% for these two scenarios. This configuration is not unlike VAX004 suggesting augmented designs could have helped inform the debate about these two hypotheses. It is clear that the performance of the designs depends dramatically on specific scenarios. In practice, careful analysis of

**Table 4**

*Simulated power for Wald and t-test of* $\mathrm{H}_0^2$ *and* $\mathrm{H}_0^3$ *under various augmented designs. The Wald tests for the* $X_0$ *design is when the actual* $X_0$ *is used in* (1) *and thus serves as an unattainable benchmark. The t-test compares the* $X_{Ci}$*'s from the placebo uninfecteds to the* $X_{0i}$ *of the vaccine* (*vaccine uninfecteds*) *to test* $\mathrm{H}_0^2$ ($\mathrm{H}_0^3$). *Standard errors for Wald tests are based on a bootstrap standard error with* 100 *bootstrap resamples. Power exceeding* 80% *is bolded. Each line is based on* 1000 *simulated vaccine trials.*

| | | | Test of $\mathrm{H}_0^2$ or $\beta_2 = 0$ | | | | | Test of $\mathrm{H}_0^3$ or $\beta_3 = 0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *t*-test | Wald tests | | | | *t*-test | Wald tests | | | |
| $n$ | $\rho$ | Scenario | CPV | CPV | BIV | CPV + BIV | $X_0$ | CPV | CPV | BIV | CPV + BIV | $X_0$ |
| 1000 | 0.25 | A | 0.34 | 0.40 | 0.01 | 0.58 | **1.0** | 0.06 | 0.04 | 0.03 | 0.04 | 0.05 |
| | 0.50 | A | 0.35 | 0.40 | **0.86** | **0.91** | **1.0** | 0.06 | 0.05 | 0.03 | 0.04 | 0.05 |
| | 0.25 | C | 0.07 | 0.08 | 0.00 | 0.06 | 0.04 | 0.23 | 0.30 | 0.13 | 0.43 | **0.98** |
| | 0.50 | C | 0.06 | 0.06 | 0.04 | 0.04 | 0.05 | 0.24 | 0.30 | 0.78 | 0.78 | **0.99** |
| | 0.25 | B | 0.19 | 0.22 | 0.00 | 0.32 | **0.99** | 0.09 | 0.17 | 0.11 | 0.22 | 0.66 |
| | 0.50 | B | 0.20 | 0.23 | 0.57 | 0.64 | **0.99** | 0.08 | 0.16 | 0.35 | 0.40 | 0.71 |
| 2500 | 0.25 | A | 0.70 | 0.74 | 0.38 | **0.91** | **1.0** | 0.07 | 0.05 | 0.03 | 0.05 | 0.05 |
| | 0.50 | A | 0.68 | 0.74 | **1.0** | **0.99** | **1.0** | 0.07 | 0.05 | 0.05 | 0.06 | 0.05 |
| | 0.25 | C | 0.06 | 0.07 | 0.02 | 0.05 | 0.040 | 0.47 | 0.52 | 0.63 | 0.78 | **1.0** |
| | 0.50 | C | 0.05 | 0.07 | 0.05 | 0.06 | 0.050 | 0.48 | 0.52 | **0.99** | **0.99** | **1.0** |
| | 0.25 | B | 0.39 | 0.44 | 0.22 | 0.64 | **1.0** | 0.12 | 0.31 | 0.30 | 0.46 | **0.97** |
| | 0.50 | B | 0.43 | 0.48 | **0.95** | **0.97** | **1.0** | 0.11 | 0.28 | 0.65 | 0.71 | **0.97** |

performance would be required to settle on a specific augmented design.

We note that a correlation of close to 0.5 may be a realistic aspiration. In the VAX004 trial, the vaccine consisted of two strains of viral gp-120, which is a sequence of 120 amino acids that comprise the outer envelope of the virus. The two strains were denoted MN and GNE8. Two nonoverlapping regions of the envelope, prone to mutations, are called the V2 and V3 loops. The amino acid sequences for the V2 loop and the V3 loop of the gp120 are completely different and thus the immune response induced by these two different loops should behave like responses to irrelevant vaccinations. Correlations between these loops were 0.42 and 0.44, respectively, for the MN and GNE8 strains. Correlations across strains were 0.34 and 0.48 (Figure 3 of Gilbert et al., 2005).

## 5. Elaborations

The methods of this article can help decide whether an improved vaccine is worth evaluating in a phase III trial. Suppose that after tinkering with the old vaccine, a new version was created, which shifted the distribution of the immune response to the right by $\Delta$. So in an obvious notation, we have $X_0(1)^{\text{old}}$ with moments $\mu_x$ and $\sigma_x^2$, while $X_0(1)^{\text{new}}$ has moments $\mu_x + \Delta$ and $\sigma_x^2$. We assume that a person with response $x$ under the old vaccine is infected with probability

$$\Phi\{\beta_0 + \beta_1 + \beta_2 x + \beta_3(x + \Delta)\}$$

under the new vaccine. Note that $\Delta$ is missing from $\beta_2$ as only $\beta_3$ reflects the causative effect of immune response. Overall, we calculate the expected event rate with the new vaccine

$$\theta_1^{\text{new}} = \int \Phi\{\beta_0 + \beta_1 + \beta_2 x + \beta_3(x + \Delta)\}\phi\big(x \bigm| \mu_x, \sigma_x^2\big)\, dx.$$

Based on the data from the trial of the old vaccine, one can estimate $\theta_0$ and $\theta_1^{\text{new}}$ and then estimate the sample size required for a phase III trial of the new vaccine with improved immune response $\Delta$. Or one might conclude that $\theta_1^{\text{new}}$ is too modest to proceed.

Closeout placebo vaccination requires time constancy of immune response. One way to examine this assumption would be to close out some fraction of the placebo uninfecteds midway through the trial, vaccinate them, and obtain their immune response, say $X_{C/2}$. Equality of the distributions of $X_{C/2}$ and $X_C$ supports time constancy of immune response provided the effect of $X_0(1)$ on $Y$ does not vary with time. To formalize this, let $Y_{C/2}$, $Y_C$ be the infection indicators over half the trial and the entire trial, respectively. If

$$p_0\{Y_{C/2} = 1 \mid X_0(1)\} \propto p_0\{Y_C = 1 \mid X_0(1)\} \tag{6}$$

then

$$\mathrm{H}_0^{T2} : X_{C/2} \overset{D}{=} X_C$$

is consistent with time constancy of immune response on an individual level. Note that if (6) does not hold, there is no point in examining $\mathrm{H}_0^{T2}$.

Testing $\mathrm{H}_0^{T2}$ need not be very costly. Simple power calculations show that for a 8800 person trial with 90% power to detect a 10% versus 8% difference in infection rates, removing 10% of placebo uninfecteds halfway through would retain at least 88% power. Additionally, comparing $X_{C/2}$ in 440 "halfway" placebo uninfecteds to $X_C$ in say the 3520 final placebo uninfecteds would give 97% power to detect a standardized difference (mean difference over standard deviation) of 0.20.

Another way to examine time constancy of immune response is to see whether the relationships between $W_0$, $X_0(1)$ and $W_0$, $X_C$ are the same in the two arms. But this also requires assumptions. For example, if the following probit model holds

$$P\{Y = 1 \mid W_0, X_0(1), Z\} = \Phi\{\beta_0 + \beta_1 Z + \beta_2 X_0(1)\}, \tag{7}$$

then

$$\mathrm{H}_0^{TW}: \{X_0(1), W_0 \,|\, Y=0, Z=1\} \overset{D}{=} (X_C, W_0 \,|\, Y=0, Z=0)$$

is consistent with time constancy of immune response. Note that $\mathrm{H}_0^{TW}$ can be tested using data readily available from a CPV trial and does not require a partial closeout halfway through the trial.

Model (1) assumes that there is no effect of $W_0$ on infection risk once $X_0(1)$ is in the model. One can specify generalizations to (1) that include $W_0$ as an additional main effect, or even allow for interaction with treatment,

$$P\{Y=1 \,|\, X_0(1), Z, W_0\} = \Phi\{\beta_0 + \beta_1 Z + \beta_2 X_0(1) + \beta_3 W_0 + \beta_4 Z X_0(1) + \beta_5 Z W_0\}, \qquad (8)$$

and likelihood construction for this model would parallel construction based on (1). It is perhaps surprising that even for our setting, where $X_0(1)$ is missing in the placebo group, this model with two interactions can be estimated provided CPV is performed. If CPV is not done, (8) is identifiable provided, for example, $\beta_5 = 0$. With CPV one could test whether $\beta_5$ and/or $\beta_3$ were 0. However, such tests would likely have poor power, as trials are powered for a treatment main effect and estimating two interactions may be difficult.

In principle, $W_0$ could be any baseline variable correlated with $X_0(1)$ and a baseline irrelevant vaccination need not be performed. Presumably, however, $W_0$ based on BIV should have a much stronger relationship with $X_0(1)$ than a variable such as race, gender, or age. An additional issue with nonimmunologically based $W_0$ is the perhaps greater concern that $\beta_3$ and or $\beta_5$ in (8) might not be zero. It is important to realize that if (8) holds with $(\beta_3, \beta_5) \neq (0, 0)$ then inference derived from fitting (the incorrect) model (1) would be misleading.

We made a simplifying assumption that there were no infections in either group until $X_0(1)$ was measured. If infections do occur over the interval [0, $m$] we can still obtain consistent estimates of the parameters provided we derive a likelihood under more assumptions. We illustrate one way. Consider a BIV design. Because the likelihoods in Section 2 factor $L(\boldsymbol{\beta}) = L_v(\beta_0 + \beta_1, \beta_2 + \beta_3) L_p(\beta_0, \beta_2)$ we can estimate $\beta_0$, $\beta_2$ using $L_p(\,)$, given consistent estimates of $\boldsymbol{\theta} = (\mu_x, \mu_z, \sigma_x^2, \sigma_z^2, \rho)$ (recall that $L_p$ depends implicitly on $\boldsymbol{\theta}$). For the vaccine group, let $\mathcal{V}(m)$ be the set of vaccinees who become infected over the interval [0, $m$] and $\mathcal{V}(R)$ be the rest of the vaccinees. Then under assumption (6) applied over [0, $m$], the likelihood for the vaccine group is proportional to

$$\left[\prod_{i \in \mathcal{V}(R)} p_1(x_{0i})^{y_i} \{1 - p_1(x_{0i})\}^{1-y_i} \phi(x_{0i}, w_{0i}; \boldsymbol{\theta})\right] \times \left\{\prod_{i \in \mathcal{V}(m)} \int p_1(u)^{y_i} \phi(u, w_{0i}; \boldsymbol{\theta})\, du\right\},$$

where $\boldsymbol{\phi}$ is the bivariate normal density function.

## 6. Final Comments

While this article has focused on immune response to an HIV vaccine, it is clear that the methods would apply to any vaccine trial. Chan et al. (2003) describe the role of immune response in vaccine development and point out the difficulty of establishing immune response as a surrogate for protection or disease burden as immune response is only measured in the vaccine group. The designs of this article allow one to use the principal surrogacy approach of Frangakis and Rubin (2002).

This article has focused on evaluating the effect of immune response on preventing infections. Current thinking on HIV vaccines is that they may have their major effect on post-infection outcomes, such as the viral load setpoint, the steady-state amount of virus in the bloodstream shortly after infection. The approach of this article could also be applied to post-infection endpoints, though this is necessarily more assumption dependent as the infected groups are not balanced by virtue of randomization (see Gilbert et al., 2003; Hudgens, Hoering, and Self, 2003).

The simulations show the profound dependence of these methods on $\rho$. Fortunately, $\rho$ can be estimated well before closeout. However, even if $\rho$ is large, there is some benefit in obtaining some CPV data as they provide a check of the imputation-based $W_0$ alone. Additionally, if it turns out to be an unanticipated immune response to the HIV vaccine, say $X_0(1)^u$ is strongly associated with infections and $W_0$ is independent of $X_0(1)^u$, a BIV-alone design would have been a mistake. CPV offers protection against this possibility. Finally, a simple $t$-test based on CPV data alone is appealing for its simplicity and transparency. Of course, CPV requires the strong assumption of time constancy of immune response.

In practice, several vaccinations over several months may be necessary during which time infections might accrue and the immune responses might wax and wane in conjunction with the vaccinations so thought is required to choose a precise time to measure $X_0(1)$. Another approach would be to develop methods that explicitly model the time-varying nature of $X_0(1)$ and use time to infection as the outcome rather than a binary indicator of infection.

Implementation of these designs could be done in an incremental fashion. Initially, small studies could be conducted to establish the extent of correlation between $W_0$ and $X_0(1)$, which irrelevant vaccine was most useful, and whether time constancy of immune response were plausible. If promising, an adaptive augmented phase III design could then be initiated.

## Acknowledgements

I am grateful to Peter Gilbert, Jorge Flores, Michael Hudgens, Michael Fay, and an associate editor for providing helpful comments on drafts of this article.

## References

Carey, V. J., Baker, C. J., and Platt, R. (2001). Bayesian inference on protective antibody levels using case-control data. *Biometrics* **57,** 135–142.

Chan, I., Li, S., Matthews, H., Chan, C., Vessey, R., Sadoff, J., and Heyse, J. (2002). Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine* **21,** 3411–3430.

Chan, I., Wang, W., and Heyse, J. (2003). Vaccine clinical trials. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd edition, S.-C. Chow (ed), 1005–1022. New York: Marcel Dekker.

Follmann, D. (2000). On the effect of treatment among treatment compliers: An analysis of the Multiple Risk Factor Intervention Trial. *Journal of the American Statistical Association* **95,** 1101–1109.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Gilbert, P., Bosch, R., and Hudgens, M. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59,** 531–541.

Gilbert, P., Peterson, M., Follmann, D., et al. (2005). Correlation between immunologic responses to rgp120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases* **191,** 666–677.

Graham, B. and Mascola, J. (2005). Lessons from failure—Preparing for future HIV-1 vaccine efficacy trials. *Journal of Infectious Diseases* **191,** 647–649.

Halloran, M. E. (1998). Vaccine studies. In *Encyclopedia of Biostatistics*, Volume 6, P. Armitage and T. Colton (eds), 4687–4694. New York: Wiley.

Halloran, M. E. and Struchiner, C. (1995). Causal inference in infectious diseases. *Epidemiology* **6,** 142–151.

Hallstrom, A. P., McAnulty, J. H., Wilkoff, B. L., Follmann, D., Raitt, M. H., Carlson, M. D., Gillis, A. M., Shih, H. T., Powell, J. L., Duff, H., and Halperin, B. D. (2001). Patients at lower risk of arrhythmia recurrence: A subgroup in whom implantable defibrillators may not offer benefit. *Journal of the American College of Cardiology* **37,** 1093–1099.

Hudgens, M. and Halloran, M. E. (2004). *Causal vaccine effects on binary post-infection outcomes.* Technical Report 04-03, Department of Biostatistics, Emory University, Atlanta, Georgia.

Hudgens, M., Hoering, A., and Self, S. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* **22,** 2281–2298.

Lachenbruch, P. A., Horne, D. A., Lynch, C. J., Tiwari, J., and Ellenberg, S. (2000). Biologics. In *Encyclopedia of Biopharmaceutical Statistics*, S.-C. Chow (ed), 47–54. New York: Marcel Dekker.

Nabel, G. (2001). Challenges and opportunities for development of an HIV vaccine. *Nature* **410,** 1002–1007.

Plikaytis, B. and Carlone, G. (2005). Statistical considerations for vaccine immunogenicity trials. Part 2: Noninferiority and other statistical approaches to vaccine evaluation. *Vaccine* **23,** 1606–1614.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* **8,** 431–440.

The rgp120 HIV Vaccine Study Group. (2005). Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *Journal of Infectious Diseases* **191,** 654–665.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* **2,** 1–26.

Rubin, D. (1978). Bayesian inference for causal effects. *Annals of Statistics* **6,** 34–58.

*Received April* 2005. *Revised November* 2005.
*Accepted January* 2006.