Published in final edited form as:

Stat Med. 2017 November 10; 36(25): 4071–4080. doi:10.1002/sim.7419.

A Method to Address Between-Subject Heterogeneity for Identification of Principal Surrogate Markers in Repeated Low-Dose Challenge HIV Vaccine Studies

Andrew J. Spieker¹ and Ying Huang²

¹University of Pennsylvania, Department of Biostatistics, Epidemiology, and Informatics, Philadelphia, PA

²Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA

Abstract

Repeated low-dose challenge designs in non-human primate studies have recently received attention in the literature as a means of evaluating vaccines for HIV prevention and identifying immune surrogates for their protective effects. Existing methods for surrogate identification in this type of study design rely on the assumption of homogeneity across subjects (namely, independent infection risks after each challenge within each subject, conditional on covariates). In practice, random variation across subjects is likely to occur due to unmeasured biologic factors. Failure to account for this heterogeneity or within-subject correlation can result in biased inference regarding the surrogate value of immune biomarkers and under-powered study designs for detecting surrogate endpoints. In this paper, we adopt a discrete-time survival model with random effects to account for between-subject heterogeneity, and we develop estimators and testing procedures for evaluating principal surrogacy of immune biomarkers. Simulation studies reveal that the heterogeneous model achieves substantial bias reduction compared to the homogeneous model, with little cost of efficiency. We recommend the use of this heterogeneous model as a complementary tool to existing methods when designing and analyzing repeated low-dose challenge studies for evaluating surrogate endpoints.

Keywords

Discrete-time survival; Heterogeneous infection risk; Random effects; Repeated low-dose challenge design

1 Introduction

Vaccine development remains an important objective in HIV/AIDS research. An effective vaccine for protection against HIV infection would be a major advance in public health research with a potential to control the HIV pandemic. Proper measures of clinical effectiveness such as time-to-event as measured in efficacy trials are crucial to understanding a vaccine's real-world impact. However, they can be time-consuming and costly to obtain for the purposes of preliminary research. Vaccine elicited immune responses measured in phase I–II studies, in contrast, have been commonly used to rank and screen out ineffective vaccine candidates in a timely fashion. It is therefore essential in HIV vaccine research to identify

good immune response biomarkers that can reliably predict a vaccine's effect in preventing HIV infection (namely, surrogate endpoints).

Among various frameworks proposed for evaluating surrogate endpoints in general biomedical research settings [1], the principal surrogate framework has been deemed particularly suitable for the motivating application to HIV [2–3]. This framework assesses causal associations based on a single trial, conditional on individuals' potential biomarker values given different treatment assignments [2]. In a typical HIV vaccine trial, only subjects not having been previously infected with similar pathogens are enrolled, such that the immune response values for subjects assigned to the placebo group are all zero. This simplifies the application of the principal surrogate framework, since only potential biomarker values assigned to the active treatment (i.e. the vaccine-induced immune response) need to be considered. This fact will become more apparent in the following section.

Research has been conducted in recent years to develop methods for evaluating vaccine-induced immune responses as principal surrogates of protection from infection. Most focus on human studies [4–7]; a few focus on non-human primate (NHP) studies in the setting of a repeated low-dose (RLD) challenge design [8–10]. RLD challenge designs have recently gained popularity as a means of identifying effective vaccines against HIV in NHP studies [11]. Under such designs, animals are randomized to either a placebo or an active vaccine, and are repeatedly exposed to small doses of the simian immunodeficiency virus (SIV); the number of challenges to infection is measured as the outcome of interest. The advantage to using the RLD design over the single high-dose design is that it reflects the low probability of HIV transmission in a human population more realistically and can potentially provide more statistical power to detect the effect of the HIV vaccine [12].

Methodology has been developed to identify surrogates of protection in the setting of the RLD design [8, 10]. All existing methodology assumes a homogeneous risk of infection conditional on vaccine-induced immune response values, treatment assignments, and observed baseline covariates, such that the probability of infection after each challenge can be expressed as a product of independent Bernoulli trials. Despite efforts to remove undue variation between subjects, however, random variation has sometimes been observed in previous NHP studies, leading to within-subject correlation. Though Regoes [13] suggests that there is typically insufficient evidence for heterogeneity in susceptibility in such studies, he notes the work of Letvin *et al.* as an example in which susceptibility depends upon the TRIM5 allele carried by the NHP [13–14]. We might expect further variation, attributable to biological factors such as inflammation, cervical ectopy, or other unmeasured covariates. Ordinarily, individual studies may be insufficiently powered to reliably rule out heterogeneity; hence, accounting for within-subject correlation would be of value if doing so is not a detriment to efficiency of estimation. In vaccine efficacy studies, it has been demonstrated that failure to account for violations to the homogeneity assumption can result in biased inference and under-powered study design [8, 15]. The impact of ignoring between-subject heterogeneity, however, has not been studied to address the other objective deemed important in an RLD design—namely, the evaluation of immune biomarkers for their surrogate potential.

The remainder of this paper is organized as follows. In Section 2, we will develop methods to account for heterogeneous risk when evaluating principal surrogate endpoints in the RLD design based on a discrete-time survival model. We will focus specifically on methods to evaluate and formally test surrogate value. In Section 3, we will elucidate the advantages of the proposed method when heterogeneity is present, comparing it to models based on the assumption of homogeneous infection risk. Bias, coverage, and efficiency will be given considerable attention, although we will particularly emphasize results surrounding power of RLD studies under various levels of heterogeneity and surrogate value. Finally, we conclude this paper with a discussion of our findings and potential future research directions.

2 Methods

In this section, we develop methods to quantify and assess the surrogate value of an immune response in the setting of an RLD design. We assume a sample of NHPs are randomized to receive either the placebo or the active vaccine; following vaccination, each NHP is repeatedly exposed in fixed time intervals (e.g., weekly) to SIV, after which each is assessed for signs of systemic infection.

Let $i = 1, \dots, N$ index study subjects, each of which is assigned to either placebo ($Z_i = 0$) or vaccine ($Z_i = 1$). Each subject receives SIV challenges repeatedly until infection or a pre-determined maximum number of challenges (denoted by $C_{\max}$), whichever comes first. Let $T_i(0)$ and $T_i(1)$ denote the potential number of challenges until infection under the placebo or active vaccine, respectively. The observed time to infection (or censoring) is then given by $T_i = \min\{T_i(0)(1 - Z_i) + T_i(1)Z_i, C_{\max}\}$. Let $\delta_i = 1(T_i(0)(1 - Z_i) + T_i(1)Z_i = C_{\max})$ denote the indicator of infection by the end of the study.

We consider an SIV-specific immune response, measured on a continuous scale and acquired from each subject shortly after they receive the vaccine or placebo. Let $S(Z)$ denote the potential immune response values (our potential surrogate) under treatment assignment Z , and S the observed value. We make the following standard assumptions: (i) the stable unit treatment value assumption (SUTVA), (ii) consistency, and (iii) ignorable treatment assignment. The first assumption states that the potential outcomes ($T_i(0)$, $T_i(1)$, $S_i(0)$, $S_i(1)$) of a subject are independent of the treatment assignment of other study subjects; the second assumption states that the observed outcome pair (S , T) is equal to the pair of potential outcomes under the treatment assignment actually received. The third assumption can be written as $(T_i(0), T_i(1), S_i(0), S_i(1)) \perp Z_i$. In an experimental RLD study of NHPs, the first condition can be ensured by not allowing interaction between the NHPs, and the third can be satisfied with proper randomization. We further assume the “constant biomarker assumption” as per Hudgens and Gilbert (namely, that $S_i(0) = 0$ for each i) [8]. This assumption holds in the RLD study design since the subjects included in the study are not exposed to SIV or similar pathogens beforehand. Without loss of generality, we presume that higher values of $S(1)$ indicate greater immune response and that $S(1) \geq 0$.

Finally, let W denote a baseline correlate of $S(1)$, measured in all study subjects (e.g., gender, age, or lab assay). This baseline covariate will be of use in obtaining information about the fundamentally unobservable (yet important) values of $S(1)$ for subjects in the

placebo group. Follmann [18] suggests that such a W could include a vaccine-elicited immune response for some other infection (e.g., rabies).

2.1 A discrete-time survival model conditional on the principal surrogate

We propose a discrete-time survival model with random effects to account for heterogeneity between subjects. Define γ_i to be a subject-specific random effect, with $\exp(\gamma_i) \sim \text{Gamma}(k = 1/\nu^2, \theta = \nu^2)$ having mean 1 and variance ν^2 . We model the per-challenge risk of infection, conditional on the random effect γ , vaccine-induced immune response $S(1)$, treatment assignment Z , and baseline predictor W , with a generalized linear mixed effects model using the complementary log-log (clog-log) link function as follows:

$$\lambda_i(Z_i, S_i(1), W_i, \gamma_i; \beta, \nu) = 1 - \exp(-\exp(\beta_0 + \beta_1 Z_i + \beta_2 S_i(1) + \beta_3 Z_i S_i(1) + \beta_4 W_i + \gamma_i)). \quad (1)$$

Conditional on Z_i , $S_i(1)$, W_i , and γ_i , T_i has a Geometric distribution with mean $1/\lambda_i$. This proposed model possesses a “memoryless” property, under which the challenge history is presumed not to influence infection risk conditional on the random effect and covariates. Regoes explores this memoryless assumption further in his work and finds evidence in its support [13]. In this section, we choose to include W as a covariate in the risk model. In practice, we may either include W or exclude it, as its primary function is to provide information about $S(1)$. We will consider both scenarios in the empirical studies of Section 3.

Note that we can express the variability of the random effect, γ , more intuitively through an intra-subject correlation term between a continuous response underlying the binary infection event. The event of an infection at challenge t for subject i , $E_i(t)$, conditional on not being infected at the previous time point, can be written in terms of a latent continuous variable $U_i(t)$ so that $E_i(t) = 1(U_i(t) > 0)$. This can be seen by letting $U_i(t) = \beta_0 + \beta_1 Z_i + \beta_2 S_i(1) + \beta_3 Z_i S_i(1) + \beta_4 W_i + \varepsilon_i$, where $-\varepsilon_i$ has a reverse extreme value distribution. Under the setup of the clog-log link with a Gamma-distributed random effect, there is a direct correspondence between the random effect variance ν^2 and the intra-subject correlation (ρ), given by $\rho = \nu^2/(\nu^2 + \pi^2/6)$ [16–17].

For the purposes of comparison in simulation studies that will follow, we will compare this model to a model that presumes homogeneous infection risk. This is analogous to (1), but without a random effect term:

$$\lambda_i(Z_i, S_i(1), W_i; \beta) = 1 - \exp(-\exp(\beta_0 + \beta_1 Z_i + \beta_2 S_i(1) + \beta_3 Z_i S_i(1) + \beta_4 W_i)). \quad (2)$$

The homogeneous risk model assumes that the memoryless property holds conditional on fixed covariates only. Note that β as given by the heterogeneous risk model (1) is a different parameter compared to β in the homogeneous model (2), due to non-collapsibility of parameters under the clog-log link.

2.2 Estimation of the Risk Model Parameters

The unobservable values of $S(1)$ for placebo-assigned subjects pose a challenge for estimating model parameters. Previously, Follmann has proposed solving this issue either through utilization of a baseline covariate W correlated with $S(1)$, or augmentation of study design [18]. We specifically expand upon the former approach in this paper.

Let $O_i \equiv (Z_i, T_i, \delta_i, S_i, W_i)$ denote the observed data vector for subject i . Suppose we observe N independent and identically distributed (i.i.d.) samples of O_i . Under the homogeneous risk model, the contribution of a subject in the vaccine arm to the likelihood conditional on covariates is given by $(1-\lambda_i)^{T_i-1}\lambda_i$ if infected by the end of the study and $(1-\lambda_i)^{T_i}$ otherwise. For a subject in the placebo arm (whose $S(1)$ is therefore unobservable), we may construct a marginal likelihood conditional on W and Z only, integrating with respect to $F_{S|W}$, the cumulative distribution function of $S(1)|W$. The contribution of a placebo subject to the likelihood would be given by $\int (1-\lambda_i)^{T_i-1}\lambda_i dF_{S|W}$ for subjects infected by the end of the study and $\int (1-\lambda_i)^{T_i} dF_{S|W}$ otherwise. These subject-specific likelihoods can be aggregated to form a full marginal likelihood.

When utilizing the heterogeneous risk model (1) for λ_i , the marginal likelihood based on all subjects in the RLD study can be obtained by further integrating each of these expressions over the presumed distribution of the random effect, γ . Letting F_Γ denote the cumulative distribution function of γ , the likelihood based on the heterogeneous model is given by

$$\begin{aligned} \mathcal{L}(\beta, \nu | O) = & \prod_{Z_i=1} \left\{ \int (1-\lambda_i)^{T_i-1} \lambda_i dF_\Gamma \right\}^{\delta_i} \left\{ \int (1-\lambda_i)^{T_i} dF_\Gamma \right\}^{1-\delta_i} \\ & \times \prod_{Z_i=0} \left\{ \int \int (1-\lambda_i)^{T_i-1} \lambda_i dF_\Gamma dF_{S|W} \right\}^{\delta_i} \left\{ \int \int (1-\lambda_i)^{T_i} dF_\Gamma dF_{S|W} \right\}^{1-\delta_i}. \end{aligned} \quad (3)$$

Use of the clog-log link together with the Gamma-distributed random effect yields a closed-form expression for the integral with respect to F_Γ . The relative advantages of this parameterization have been documented elsewhere [15, 17, 19–20]. The derivation of this integral is presented in Appendix A from the Supplementary Materials, and its final form is given as follows:

$$\begin{aligned} \mathcal{L}(\beta, \nu | O) = & \prod_{Z_i=1} \{I_{\beta, \nu}(T_i-1, Z_i, S_i, W_i) - I_{\beta, \nu}(T_i, Z_i, S_i, W_i)\}^{\delta_i} \{I_{\beta, \nu}(T_i, Z_i, S_i, W_i)\}^{1-\delta_i} \\ & \times \prod_{Z_i=0} \left\{ \int (I_{\beta, \nu}(T_i-1, Z_i, S_i, W_i) - I_{\beta, \nu}(T_i, Z_i, S_i, W_i)) dF_{S|W} \right\}^{\delta_i} \\ & \times \left\{ \int I_{\beta, \nu}(T_i, Z_i, S_i, W_i) dF_{S|W} \right\}^{1-\delta_i}, \end{aligned} \quad (4)$$

where $I_{\beta, \nu}(T_i, Z_i, S_i, W_i) = [T_i \exp(\beta_0 + \beta_1 Z_i + \beta_2 S_i(1) + \beta_3 Z_i S_i(1) + \beta_4 W_i) \nu^2 + 1]^{-1/\nu^2}$.

Integration with respect to $F_{S|W}$ in the placebo group must be addressed in order to estimate β and ν regardless of whether the homogeneous risk or heterogeneous risk approach is used. Maximum estimated likelihood estimators (MELEs), developed by Pepe and Fleming, have proven useful in the homogeneous risk model, and so we choose to employ it here [8, 21]. Under this methodology, $F_{S|W}$ in (4) is estimated by $\hat{F}_{S|W}$, from which we may maximize $\log \mathcal{L}(\beta, \nu; \mathbf{O}; \hat{F}_{S|W})$ to obtain MELEs $\hat{\beta}$ and $\hat{\nu}$. These estimates are consistent for β and ν . We assume without loss of generality that $\nu > 0$ in order to bypass the need to use constrained optimization techniques and to avoid boundary convergence challenges.

2.3 Testing Surrogate Value

In this section, we describe a procedure for testing surrogate value. Note that more negative values of β_3 indicate greater surrogate value of S . Hence, a simple way to establish surrogate value is to test the null hypothesis $H_0: \beta_3 = 0$ against the alternative hypothesis $H_1: \beta_3 < 0$. We propose use of a one-sided level- α parametric bootstrap testing procedure for testing H_0 against H_1 [8, 22]. In this testing procedure, a restricted MELE is first fit subject to the constraint that $\beta_3 = 0$; B simulation replicates are then generated under this null model, from each of which constrained estimates of β_3 are obtained. Letting $\mathcal{B} = \{\hat{\beta}_3^b\}_{b=1}^B$ denote this set, the unrestricted MELE $\hat{\beta}_3$ is then compared to the α -quantile of the set $\mathcal{B}$, rejecting H_0 if $\hat{\beta}_3$ is less than the α -quantile.

Prior work made use of the proportion associative effect (PAE) to measure surrogate value [8]. This can be summarized as follows: for some contrast function $h(x, y)$, the expected associative effect (EAE) is defined to be $\mathbb{E}[h(\mathbb{E}[T_X(0)|S_X(1)], \mathbb{E}[T_X(1)|S_X(1)])|S_X(1) > S_X(0)]$, and the expected dissociative effect (EDE) is given by $\mathbb{E}[h(\mathbb{E}[T_X(0)|S_X(1)], \mathbb{E}[T_X(1)|S_X(1)])|S_X(1) = S_X(0)]$. The PAE is then given by $|\text{EAE}|/(|\text{EAE}| + |\text{EDE}|)$. Under the clog-log link function, for example, the choice $h(x, y) = \log(-\log(1 - 1/x)) - \log(-\log(1 - 1/y))$ leads to the following expression for the PAE:

$$\text{PAE} = \frac{|\beta_1 + \kappa\beta_3|}{|\beta_1| + |\beta_1 + \kappa\beta_3|}, \quad (5)$$

where $\kappa = \mathbb{E}[S_X(1)|S_X(1) > 0]$, and β may be estimated from the homogeneous model described in Section 2.1. Values of the PAE between 0 and 0.5 signify lack of surrogate value, and values between 0.5 and 1 indicate increasing surrogate value. In the heterogeneous risk setting, one may define a PAE-like measure as a function of model coefficients. However, given the lack of a clear clinical interpretation of the PAE, and the lack of invariance under choices of h , we choose to conduct the simpler tests on the basis of β_3 , and quantify surrogate value on the basis of the vaccine efficacy curve, described in Section 2.4.

2.4 Quantifying Surrogate Value Using Vaccine Efficacy Curves

In this section, we consider estimation of the per-challenge vaccine efficacy curve as a function of the principal surrogate $S(1)$ in order to quantify surrogate value. Here, vaccine efficacy (VE) is represented as a function of $S(1)$, defined at each $S(1)$ value to be the

percent reduction in infection rate, comparing the vaccine recipients with observed immune response $S(1)$ and baseline covariate W to the placebo recipients with counterfactual immune response $S(1)$ and baseline covariate W . The vaccine efficacy curve as a function of $S(1)$ has been proposed as a way to quantify and compare surrogate values [4, 7]; this measure provides insight into the range of vaccine efficacies achievable at corresponding vaccine-induced immune response levels. We are interested in identifying surrogates with larger variability in $VE(S(1))$ such that vaccine efficacy increases with $S(1)$. Such a finding would suggest a potential to achieve a large vaccine efficacy by developing vaccine that generates a high level of immune response.

Under the heterogeneous risk model (1), the per-challenge marginal infection risk given assignment $Z_i = z$ and baseline correlate $W_i = w$ is given by $\int \lambda(z, S(1), w, \gamma) dF_\Gamma$. The per-challenge vaccine efficacy curve is hence defined as

$$VE(s|w) = 1 - \frac{\int \lambda(Z=1, s, w, \gamma) dF_\Gamma}{\int \lambda(Z=0, s, w, \gamma) dF_\Gamma}. \quad (6)$$

As shown in Appendix B of the Supplementary Materials, we have that $\int \lambda_i dF_\Gamma = 1 - I_{\beta, \nu}(T_i = 1; Z_i, S_i, W_i)$. Hence, the vaccine efficacy curve can be represented as a closed-form expression of the heterogeneous risk model parameters:

$$VE(s|w) = 1 - \frac{1 - I_{\beta, \nu}(T=1, Z=1, s, w)}{1 - I_{\beta, \nu}(T=1, Z=0, s, w)}. \quad (7)$$

In numerical studies, we estimate the vaccine efficacy curve by substituting estimates of β and ν into (7). Standard errors for the estimated vaccine efficacy curve can be computed in a point-wise fashion using the nonparametric bootstrap, constructed at fixed values of s . Confidence intervals can also be created by using appropriate quantiles of the bootstrap replicates.

In contrast, when the homogeneous risk model (2) is fit to the data, one can directly estimate $VE^{\text{hom}}(s|w) = 1 - \lambda(Z=1, s, w; \beta) / \lambda(Z=0, s, w; \beta)$ without the need to integrate over the distribution of the random effect.

3 Simulation Study

We conduct a series of simulation studies to compare our heterogeneous model to the homogeneous model that ignores between-subject heterogeneity. For all simulations, we fix $N = 60$ subjects total (thirty per group) and set the maximum number of challenges at $C_{\max} = 30$ under the RLD design. Suppose $(S(1), W)$ is generated from a bivariate normal distribution with mean vector $\mu = (0.5, 0.5)^T$, common variance $\sigma_s^2 = \sigma_w^2 = 0.15$, and correlation $\rho = 0.5$, with left-truncation so that $S(1) \geq 0$ and $W \geq 0$. These parameters are estimated accounting for left-truncation, as described by Lyles *et al.* in order to obtain $\hat{F}_{S|W}$ [23]. Conditional on $\hat{F}_{S|W}$, the MELEs $\hat{\beta}$ and $\hat{\nu}$ can be obtained.

For simplicity, we first consider a data generation mechanism in which $T_i \perp W_i | S_i(1), Z_i$, such that $\lambda_i(Z_i, S_i(1), W_i, \gamma_i, \beta, \nu) = 1 - \exp(-\exp(\beta_0 + \beta_1 Z_i + \beta_2 S_i(1) + \beta_3 Z_i S_i(1) + \gamma_i))$. We consider four different levels of heterogeneity: none ($\nu^2 = 0$; $\rho = 0$), mild ($\nu^2 = 1/3$; $\rho \approx 0.17$), moderate ($\nu^2 = 3/4$; $\rho \approx 0.31$), and strong ($\nu^2 = 8/5$; $\rho \approx 0.49$). For all cases, β is chosen so that the expected per-challenge infection risk is 0.6 in the placebo group, with an average vaccine efficacy per-challenge of approximately 0.5. We consider three different levels of surrogate value: none ($\beta_3 = 0$), moderate ($\beta_3 = -1.8$), and high ($\beta_3 = -4$).

For each scenario, we generate one-thousand replicates fitting both the homogeneous and heterogeneous models. We obtain estimates of the per-challenge infection risks and vaccine efficacy at a range of values of s between 0 and 1, along with non-parametric 90% percentile-based bootstrap confidence intervals (500 bootstrap replicates). We evaluate the bias and empirical standard errors for estimates of $VE(s)$ and $VE^{\text{hom}}(s)$, and estimate the coverage of the bootstrap confidence intervals. We conduct the parametric bootstrap hypothesis test for β_3 as described in Section 2.3, taken at the nominal $\alpha = 0.10$ level with $B = 500$ bootstrap replicates. All simulations were conducted using R [24].

3.1 Estimation of Vaccine Efficacy Curves and Test of Surrogate Value

Table 1 summarizes the results for estimation of the vaccine efficacy curve at $S(1) = 0.5$. Presented are the estimated bias, Monte-Carlo standard error, and coverage probabilities from the bootstrap confidence intervals under the various choices for β_3 and ν^2 .

In the setting of no heterogeneity ($\nu^2 = 0$), both the homogeneous and heterogeneous models estimate the vaccine efficacy well, with approximately proper coverage and generally little bias, as expected. The cost of efficiency in using the heterogeneous model relative to the homogeneous model appears to be minimal when the homogeneity assumption holds. As between-subject heterogeneity increases, bias from the homogeneous model becomes more substantial, and bootstrap confidence intervals tend to show under-coverage. This is especially true under higher surrogate value (more negative values of β_3). In contrast, the heterogeneous model generally shows approximately correct coverage and little bias across the values of $S(1)$ considered. Of note is that bias for both the heterogeneous and homogeneous models appeared to be higher for values of $S(1)$ near the boundary ($S(1) = 0.2$ and $S(1) = 0.8$, as presented in Appendix C of the Supplementary Materials).

Figure 1 depicts the infection risk and vaccine efficacy curves for the setting of moderate surrogate value ($\beta_3 = -1.8$), under moderate heterogeneity ($\nu^2 = 3/4$). Appendix D of the Supplementary Materials presents the analogous figures in the setting of no surrogate value ($\beta_3 = 0$) and the setting of high surrogate value ($\beta_3 = -4$). The heterogeneous model estimates these curves very well, and the homogeneous model shows substantial bias for all three values of β_3 . Unsurprisingly, we found that the overall bias is smaller under lower heterogeneity, and more extreme under higher heterogeneity. Appendix E of the Supplementary Materials depict a series of figures in the setting where $T_i \not\perp W_i | S_i(1), Z_i$ (i.e., a setting in which the baseline covariate influences the infection risk beyond predicting the immune marker). In this setting, a family of vaccine efficacy curves must be considered, conditional on W . The results from this study are consistent with our main simulation study.

Also presented in Table 1 are results from the parametric bootstrap hypothesis test for surrogate value. The test of $\beta_3 > 0$ under the heterogeneous model achieves a type-I error rate fairly close to the nominal 10% level, regardless of whether or not heterogeneity is present; the type-I error rate does not appear to depend on the magnitude of heterogeneity. In contrast, the test assuming homogeneity has a substantially inflated type-I error rate even under modest heterogeneity ($\nu^2 = 1/3$), rejecting H_0 at inappropriately high rates. Further investigation revealed that the estimated level asymptotes at approximately 0.2 in this setup; this will be explored in Section 3.2.

Under moderate surrogate value ($\beta_3 = -1.8$), the test assuming homogeneity has higher power than the test allowing for heterogeneity under both $\nu^2 = 0$ and $\nu^2 = 1/3$, the latter at a cost of an inflated type-I error rate. However, under $\nu^2 = 3/4$ and $\nu^2 = 8/5$, the test using the heterogeneous model provides higher power for detecting the existence of surrogate value compared to the test that ignores heterogeneity. That is to say that while both methods see a power reduction with increased heterogeneity, the rate of power reduction for the proposed model allowing for heterogeneity is less severe than that of the homogeneous model.

Comparable patterns hold under high surrogate value ($\beta_3 = -4$). The test assuming homogeneity achieves high power (83%) when heterogeneity is absent. However, in the setting of moderate heterogeneity, the test assuming homogeneity provides an estimated 33% power compared to 55% power from the test assuming a heterogeneous model. The uniformly similar or superior power demonstrated by the test using the heterogeneous model in the presence of heterogeneity (compared to the test based on the homogeneous model) makes the heterogeneous model particularly appealing.

3.2 Power and Sample Size

In a follow-up simulation study, we utilized the same parameters as in our main simulation study, varying sample size under two levels of surrogate value (none/high) and two levels of heterogeneity (none/moderate).

Figure 2 depicts the estimated power of the parametric bootstrap test under sample sizes ranging from 10 to 50 NHPs per arm (i.e., ranging from $N = 20$ to $N = 100$ total NHPs). In the setting of high surrogate value and no heterogeneity, the power curves are approximately comparable, with the heterogeneous model providing slightly lower power. Based on our simulation setup, it appears that use of the heterogeneous model in the absence of heterogeneity translates to approximately 5 NHPs per arm in sample size cost. For example, a power of approximately 65% can be achieved with 20 NHPs per arm in the homogeneous risk model, but would require 25 NHPs per arm to achieve comparable power under the heterogeneous risk model. However, in the setting of moderate heterogeneity, the heterogeneous risk model is of far greater power than the homogeneous model, attributable to high levels of bias associated with the latter. Under moderate heterogeneity, a sample size of $N = 100$ yields under 40% power, where the heterogeneous model can achieve comparable power with a sample size of just over $N = 40$. This finding suggests that use of the homogeneous model can result in highly underpowered designs when variation exists across NHPs.

Appendix F of the Supplementary Materials shows the estimated levels in the setting of no surrogate value. When there is no heterogeneity, the tests based on the homogeneous and heterogeneous risk models approximately achieve the nominal $\alpha = 0.1$ level. When the heterogeneity is increased, the homogeneous risk model rejects H_0 too often, consistent with results from Table 1. This problem appears to worsen with increasing sample size.

4 Discussion

We have presented a discrete-time survival model with random effects for estimating and testing the principal surrogate value of immune biomarkers in the presence of between-subject heterogeneity for RLD challenge study designs. Though no RLD studies have been conducted to meet this objective to date, existing research on vaccine efficacy in RLD studies has prompted methods development to meet this objective going forward. The major goal of our study was therefore to develop an approach that would allow us to properly design a RLD study in the presence of heterogeneity when the primary objective is immune biomarker assessment. Through numerical studies, we demonstrated that comparative methods based on a homogeneous risk model can produce biased estimates and invalid inference regarding surrogate value when within-subject correlation is present. Ignoring heterogeneity when designing an RLD study for detecting surrogates can lead to under-powered study designs. Our proposed methods, on the other hand, provide valid inference for detecting surrogate value at various levels of heterogeneity studied, and interestingly can even provide more power for detecting surrogate value than the homogeneous model in settings where the intra-subject correlation is high.

When the population is truly homogeneous, there is a minor loss of power in using the heterogeneous model due to the estimation of an additional parameter. Since heterogeneity may be present in RLD studies due to unobserved covariates and biologic factors, we argue that this cost of efficiency is negligible, considering the appreciable bias reduction compared to the homogeneous model in the presence of heterogeneity.

We observed two important results surrounding the parametric bootstrap test based on the model assuming homogeneity: (1) the test has an inflated type-I error rate when heterogeneity is present; (2) although the test achieves higher power than the heterogeneous model in the absence of heterogeneity, its power decreases very quickly when the heterogeneity level increases. The former suggests that ignoring heterogeneity when it is present may lead to false discovery of immune biomarkers without surrogate value, resulting in poor recommendation of candidate vaccines for advancing to efficacy trials based on their immunogenicity in phase I–II studies. The latter suggests that sample sizes (i.e., N and $C_{\max}$) determined under the assumption of homogeneity can lead to substantially under-powered designs for identifying immune surrogates when moderate levels of heterogeneity exist. In contrast, the nominal type-I error rate is achieved under the heterogeneous model regardless of the extent of heterogeneity, with minimal power sample size cost in the absence of heterogeneity. This suggests that when designing an RLD challenge experiment for identification of immune surrogates in practice, it is important to study the heterogeneous model assuming some reasonable level of heterogeneity as a complementary tool to the homogeneous model.

Finally, we would like to comment on our choice for statistical measures of surrogate value. Previously, researchers have developed summary measures such as the PAE for characterizing surrogate value under the homogeneous model. The PAE has the advantage of being between 0 and 1 and therefore can put different markers on a common scale for comparison. Moreover, it can be represented as a simple function of risk model coefficients for certain contrast functions. However, since the magnitude of the PAE depends upon the contrast function, the particular contrast function that leads to a simple representation of the PAE may not be the contrast function of interest in clinical practice. The need to integrate the PAE over the distribution of the random effect in the presence of heterogeneity further obscures our ability to describe the clinical meaningfulness of this measure. In HIV vaccine efficacy trials, vaccine efficacy defined as the percent reduction in infection risk is of ultimate interest. We therefore choose to assess surrogate value using the vaccine efficacy curves in the RLD design, which has been proposed for characterizing surrogate markers in human trials [7].

4.1 Study Limitations and Future Directions

In this paper, we studied methods in which a baseline predictor of vaccine-induced immune response (BIP) was utilized as a way to obtain information about missing principal surrogate values. This approach has been proposed previously in human trials and RLD studies for handling continuously measured missing potential outcomes. Gilbert and Hudgens show that under the BIP design, the risk model is identifiable under some untestable model assumptions [4]. One possible strategy to obtain a good BIP is via elicitation of an immune response to another vaccine, as suggested by Follmann [18]. In RLD studies, non-HIV vaccination at baseline must be pre-planned and budgeted at the study design's conception, since NHPs will ostensibly not have received prior vaccinations the way that, for example, human subjects might. Another way to handle missing potential outcomes, proposed by Follmann, is called close-out placebo vaccination (CPV) [18]. Under this design, a portion of unaffected subjects in the placebo arm are vaccinated at the end of the standard trial period and their immune response measured as substitutes for missing potential outcomes. The advantage of the CPV design is that it allows nonparametric estimation of the potential outcome distribution, and consequently, evaluation of risk model assumptions. One could add a CPV component to the end of the RLD design. Application of our methods for modeling between-subject heterogeneity when evaluating continuous immune surrogates would be straightforward; we would still use vaccine recipients to estimate $F_{S|W}$, but subjects in the CPV component will contribute their given $S(1)$ value to the likelihood.

Recently, for evaluation of binary immune surrogates, Long *et al.* have proposed another augmented design component for RLD studies named the crossover vaccination design (CrV), in which a pre-specified proportion of uninfected placebo subjects are randomly assigned to receive vaccination after each challenge [10]. The CrV design has been developed under the homogeneous risk model. To further incorporate within-subject correlation, one can similarly integrate the likelihood over the presumed distribution of the random effect and the conditional distribution $S|W$. Unlike in the BIP and CPV setting, however, the Gamma-distributed random effect with clog-log link no longer results in a closed-form expression for the marginal likelihood. Further research is needed to develop

and evaluate methods for handling between-subject heterogeneity in the CrV design. In particular, the challenge of numerical integration over a bivariate function will need to be addressed; another complication that arises in the presence of heterogeneity is that the probability of crossing over to the vaccine group for a placebo subject varies with the underlying random effect. These problems are currently under investigation. Additional methodology research is also justified in order to account for more complex data structures in RLD study designs. For example, it would be of interest to relax the “memoryless” property of this class of models to allow challenge history to influence risk.

Evaluating principal surrogate value is a more challenging task than evaluating vaccine efficacy in RLD studies. We are unaware of any existing RLD studies that have these new design components implemented for surrogate marker evaluation. Yet, there have been increasing interests in the field to designing RLD studies that incorporate surrogate marker evaluation. Having proper statistical methods developed is important to ensure that we can design future RLD studies with adequate power for surrogate marker identification and valid inference when analyzing such studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Sources

This work was supported by the National Institutes of Health (Grants R01 GM 106177 and R37 AI 05465).

References

1. Joffe M, Greene T. Related causal frameworks for surrogate outcomes. *Biometrics*. 2009; 65:530–538. [PubMed: 18759836]
2. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
3. Gilbert P, Hudgens M, Wolfson J. Commentary on “Principal stratification – a goal or a tool?” by Judea Pearl. *The International Journal of Biostatistics*. 2011; 7 Article 36.
4. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics*. 2008; 64:1146–1154. [PubMed: 18363776]
5. Li Y, Taylor JMG, Elliott MR. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*. 2010; 66:523–531. [PubMed: 19673864]
6. Huang Y, Gilbert PB. Comparing biomarkers as principal surrogate endpoints. *Biometrics*. 2011; 67:1442–1451. [PubMed: 21517791]
7. Huang Y, Gilbert PB, Wolfson J. Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics*. 2013; 69:301–309. [PubMed: 23409839]
8. Hudgens MG, Gilbert PB. Assessing vaccine effects in repeated low-dose challenge experiments. *Biometrics*. 2009; 65:1223–1232. [PubMed: 19397589]
9. Hudgens MG, Gilbert PB, Mascola JR, Wu C, Barouch DH, Self SG. Power to Detect HIV Vaccine Effects in Repeated Low-Dose Challenge Experiments. *The Journal of Infectious Diseases*. 2009; 200:609–613. [PubMed: 19591571]
10. Long DM, Hudgens MG, Wu C. Surrogates of protection in repeated low-dose challenge experiments. *Statistics in Medicine*. 2015; 34:1747–1760. [PubMed: 25628249]

11. Qureshi H, Ma ZM, Huang Y, Hodge G, Thomas MA, DiPasquale J, DeSilva V, Fritts L, Bett AJ, Casimiro DR, Shiver JW, Robert-Guroff M, Robertson MN, McChesney MB, Gilbert PB, Miller CJ. Low-dose penile SIVmac251 exposure of rhesus macaques infected with adenovirus type 5 (Ad5) and then immunized with a replication-defective Ad5-based SIV Gag/Pol/Nef vaccine recapitulates the results of the phase IIb Step trial of a similar HIV-1 vaccine. *Journal of Virology*. 2012; 86:2239–2250. [PubMed: 22156519]
12. Ellenberger D, Otten RA, Li B, Aidoo M, Rodriguez IV, 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. *Virology*. 2006; 352:216–225. [PubMed: 16725169]
13. Regoes RR. The role of exposure history on HIV acquisition: insights from repeated low-dose challenge studies. *PLOS Computational Biology*. 2012; 8:e1002767. [PubMed: 23180981]
14. Letvin NL, Rao SS, Montefiori, Seaman MS, Sun Y, Lim SY, Yeh WW, Asman M, Gelman RS, Shen L, Whitney JB, Seoighe C, Lacerda M, Keating S, Norris PJ, Hudgens MG, Gilbert PB, Buzby AP, Mach LV, Zhang J, Balachandran H, Shaw GM, Schmidt SD, Todd JP, Dodson A, Mascola JR, Nabel GJ. Immune and genetic correlates of vaccine protection against mucosal infection by SIV in monkeys. *Science Translational Medicine*. 2011; 3:81ra36.
15. Kang C, Huang Y, Miller CJ. A discrete-time survival model with random effects for designing and analyzing repeated low-dose challenge experiments. *Biostatistics*. 2015; 16:295–310. [PubMed: 25190513]
16. Rodriguez G, Elo I. Intra-class correlation in random-effects models for binary data. *The Stata Journal*. 2003; 3:32–46.
17. Coull BA, Houseman EA, Betensky RA. A computationally tractable multivariate random effects model for clustered binary data. *Biometrika*. 2006; 93:587–599.
18. Follmann D. Augmented designs to assess immune response in vaccine trials. *Biometrics*. 2006; 62:1161–1169. [PubMed: 17156291]
19. Conway MR. A random effects model for binary data. *Biometrics*. 1990; 46:317–328.
20. Scheike TH, Jensen TK. A discrete survival model with random effects: an application to time to pregnancy. *Biometrics*. 1997; 53:318–329. [PubMed: 9147597]
21. Pepe MS, Fleming TR. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*. 1991; 86:108–113.
22. Davison, AC., Hinkley, DV. *Bootstrap Methods and their Application*. Cambridge University Press; New York: 1997.
23. Lyles RH, Williams JK, Chuachoowong R. Correlating Two Viral Load Assays with Known Detection Limits. *Biometrics*. 2001; 57:1238–1244. [PubMed: 11764265]
24. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2013. URL <http://www.R-project.org/>

Figure 1. Average infection risk (left) and median vaccine efficacy curves (right) for both the heterogeneous and homogeneous models: $\nu_2 = 3/4$ and $\beta_3 = -1.8$. Risk curves are plotted for $Z = 1$ (upper) and $Z = 0$ (lower). The truth is represented by a solid curve.

Figure 2. Power of the parametric bootstrap test for testing surrogate value as sample size varies under high surrogate value (Left: No heterogeneity; Right: Moderate heterogeneity). The dashed gray line represents the nominal $\alpha = 0.1$ level.

Simulation results for estimation of vaccine efficacy curve at $s = 0.5$ and parametric bootstrap test power. The true vaccine efficacy used for simulating the data ($VE = VE(s = 0.5)$) is provided, given as in Section 2.4. Estimated bias (averaged across one-thousand simulation replicates), Monto-Carlo standard error estimates, and empirical coverage of two-sided 90% Wald-based confidence intervals are presented. Note that as β_3 decreases, surrogate value increases, and as ν^2 increases, heterogeneity increases.

Table 1

Vaccine Efficacy Curve At $s = 0.5$										Bootstrap Test	
β_3	ν^2	Homogeneous				Heterogeneous				Homogeneous	Heterogeneous
		VE	Bias	SE	CP	VE	Bias	SE	CP	Power	Power
0.0	0	0.50	0.00	0.10	0.89	0.50	0.00	0.10	0.89	0.08	0.08
	1/3	0.49	0.08	0.13	0.85	0.49	0.01	0.13	0.90	0.20	0.12
	3/4	0.49	0.17	0.14	0.70	0.49	0.00	0.15	0.90	0.23	0.12
-1.8	8/5	0.52	0.32	0.08	0.27	0.52	0.02	0.14	0.88	0.20	0.09
	0	0.49	-0.02	0.12	0.88	0.49	-0.03	0.12	0.89	0.51	0.44
	1/3	0.48	0.07	0.14	0.85	0.48	0.01	0.14	0.91	0.65	0.39
	3/4	0.52	0.17	0.12	0.69	0.52	0.02	0.14	0.90	0.24	0.30
	8/5	0.56	0.29	0.08	0.28	0.56	0.01	0.14	0.91	0.02	0.20
	-4.0	0	0.78	-0.02	0.07	0.88	0.78	-0.02	0.08	0.88	0.76
	1/3	0.76	0.03	0.08	0.87	0.76	-0.01	0.09	0.91	0.68	0.64
	3/4	0.71	0.08	0.09	0.74	0.71	-0.01	0.12	0.90	0.33	0.55
	8/5	0.57	0.28	0.08	0.29	0.57	0.02	0.14	0.89	0.06	0.38