Research Article

Statistics in Medicine

Received 30 March 2015, Accepted 7 March 2016 Published online 31 March 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6952

Power/sample size calculations for assessing correlates of risk in clinical efficacy trials

Peter B. Gilbert,*† Holly E. Janes and Yunda Huang

In a randomized controlled clinical trial that assesses treatment efficacy, a common objective is to assess the association of a measured biomarker response endpoint with the primary study endpoint in the active treatment group, using a case-cohort, case-control, or two-phase sampling design. Methods for power and sample size calculations for such biomarker association analyses typically do not account for the level of treatment efficacy, precluding interpretation of the biomarker association results in terms of biomarker effect modification of treatment efficacy, with detriment that the power calculations may tacitly and inadvertently assume that the treatment harms some study participants. We develop power and sample size methods accounting for this issue, and the methods also account for inter-individual variability of the biomarker that is not biologically relevant (e.g., due to technical measurement error). We focus on a binary study endpoint and on a biomarker subject to measurement error that is normally distributed or categorical with two or three levels. We illustrate the methods with preventive HIV vaccine efficacy trials and include an R package implementing the methods. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: case-cohort design; case-control design; immune response biomarkers; measurement error; principal stratification; two-phase sampling design; vaccine efficacy trial

1. Introduction

Commonly, clinical efficacy trials randomize study participants to receive a treatment or control preparation (e.g., placebo) at one or more visits and follow these participants for occurrence of the primary clinical study endpoint. The primary objective assesses treatment efficacy against the clinical endpoint, and a common secondary objective assesses the association of intermediate response endpoints (e.g., biomarkers) measured after the administration of treatment with primary endpoint occurrence in the active treatment group. Applications of this secondary objective include developing prognostic biomarkers and providing information for other analysis objectives such as surrogate endpoint and mediation assessment. Typical statistical approaches for assessing such correlates of risk (CoRs) have included logistic or Cox proportional hazards regression models that account for the sampling design that was used for measuring the biomarkers (e.g., [1–4]).

For power calculations to detect CoRs in a cohort such as an active treatment group, many methods have been developed for case-cohort studies (e.g., [5]), case-control studies (e.g., [6, 7]), and the generalization of case-control studies to two-phase sampling studies (e.g., [8]). However, the available approaches typically do not account for the level of clinical treatment efficacy overall and in biomarker response subgroups, precluding interpretation of the results in terms of potential correlates of efficacy/protection. We develop an approach to CoR power/sample size calculations that accounts for this issue, which is important because if the power calculations are based solely on the biomarker-outcome association in the active treatment group, then one could design a case-control study to, say, have 90% power to detect a biomarker-outcome odds ratio of 0.5, but not realize that this power is achieved under a tacit

Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

Correspondence to: Peter B. Gilbert, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave North, PO Box 19024, Seattle, WA, U.S.A.

† *E-mail: pgilbert@fredhutch.org*

assumption that the endpoint rate is higher in the active arm than the control arm for the subgroup with lowest biomarker responses. By specifying overall treatment efficacy and biomarker-specific treatment efficacies as input parameters, our approach makes transparent in the power calculations the link between the CoR effect size in the active treatment arm and the corresponding difference in biomarker-specific treatment efficacies.

In addition, our approach accounts for the component of inter-individual variability of the biomarker that is not biologically relevant (e.g., due to technical measurement error of the device employed to measure a biological response), which is important because the degree of measurement error of the biomarker heavily influences power of the CoR analysis, such that accounting for this issue is needed to obtain accurate power calculations. In our approach, the user inputs a parameter ρ defined as the estimated fraction of the biomarker's variance that is potentially biologically relevant for protection and displays how power and sample size requirements vary with ρ.

Our approach can be used for a general binary clinical endpoint model with case-cohort, case-control, or two-phase sampling of the biomarker, using without replacement or Bernoulli sampling. We illustrate the approach with a logistic regression model and case-control without replacement sampling. For rare event studies (e.g., with cumulative endpoint rate less than 10%), we found in simulations that the power for the logistic regression model tends to be very similar to that for a Cox regression model [9]; thus, in this setting, the approach may provide sufficiently accurate power results for time-to-event CoR analysis. The simplification afforded by using a binary outcome is helpful for focusing attention on the two issues listed previously.

Related research has developed power calculators of testing procedures for assessing the association of a *true* biomarker subject to measurement error and a sub-sampling design with an outcome (e.g., [10–12]). Here, we depart from this research objective by developing a power calculator of testing procedures for assessing the association of a *measured/observed* biomarker that has components of variability thought to be not possibly associated with the outcome. Whereas the former testing procedures incorporate bias-correction techniques, leveraging, for example, validation sets or replicate biomarker measurements, our power calculator may be used with a large number of available hypothesis testing procedures from the case-cohort/case-control/two-phase sampling statistical methods literature (going back to Horvitz and Thompson [13]), where the methods do not need bias-correction techniques. Thus, the contribution of this work is to provide more interpretable and accurate power calculations for the common scientific endeavor to understand power for detecting the association of a measured/observed biomarker with the outcome. Moreover, previous work has developed power calculation formulas for associating a measured biomarker subject to measurement error with a dichotomous outcome; for example, [14–16] considered a normally distributed biomarker following a classical measurement error model, with application to logistic regression correlates analysis.

While the newly proposed power calculator applies for general randomized controlled two-group clinical trials, for definiteness, we focus on preventive vaccine efficacy trials, which randomize study participants to receive a candidate vaccine or placebo at one or more visits, and follow these participants for occurrence of clinically significant infection with the pathogen under study [17]. The primary objective assesses vaccine efficacy (*VE*) defined as the multiplicative reduction (vaccine versus placebo) in the rate of the primary endpoint, and a secondary objective assesses the association of immune response biomarkers measured shortly after vaccination with the primary endpoint. For settings where some trial participants were previously infected with the pathogen (e.g., influenza), this analysis is done for each of the vaccine and placebo groups or pooling over the groups, and for settings where trial participants have not been previously infected with the pathogen (e.g., HIV), such that the immune response biomarker does not vary in the placebo group [18], this analysis is done either pooling over the vaccine and placebo groups or in the vaccine group only. In the vaccine field, such analyses have been named CoR analyses (e.g., [19, 20]), and for definiteness, we focus on assessing a CoR in the vaccine group. The approach is illustrated with power calculations for the RV144 HIV vaccine efficacy trial after the primary analysis was conducted and with sample size calculations for the prospective design of a sequel HIV vaccine efficacy trial being planned by the HIV Vaccine Trials Network.

Section 2 describes the study set-up, parameters of interest, and identifiability assumptions. Section 3 describes the power and sample size calculation approach. Section 4 illustrates the power/sample size calculator with the two examples, and Section 5 concludes with discussion. Supporting Information Appendix A discusses how to unbiasedly characterize the biomarker distribution accounting for the sampling design, Supporting Information Appendix B provides selected mathematical details of the power calculation methods, and Supporting Information Appendix C addresses the important topic of how to estimate the noise level of the biomarker. Supporting Information Appendix D presents supplementary figures for the two illustrations and Supporting Information Appendix E summarizes how to use the R package.

2. Study set-up, parameters of interest, identifiability assumptions

2.1. *Randomized clinical trial for assessing vaccine efficacy*

We consider a double-blind clinical trial that randomizes participants to vaccine or placebo, with Z the indicator of assignment to vaccine and W baseline covariates. Let S be the immune response biomarker measured at a fixed time τ post-randomization, which we assume to be continuous or trichotomous, with the case of dichotomous S covered as a special case. Participants are followed for occurrence of the primary clinical study endpoint, clinically significant infection with the pathogen, with follow-up through time $\tau_{\max}$, with T the time from randomization until the study endpoint and $Y \equiv I[T \leqslant \tau_{\max}]$ the binary outcome of interest. Let $Y^{\tau} \equiv I[T \leqslant \tau]$ and V^{τ} be the indicator that a subject attends the visit at τ. Fitting to the motivating application, we focus on settings where it is only interesting to study the association of S with Y for subjects who did not experience the event before the biomarker is measured. Therefore, subjects with $(1-Y^{\tau})V^{\tau}=1$ are the subgroup observed to be at-risk at τ who could potentially have S measured for the association study.

Because S is expensive to measure, a case-cohort, case-control, or two-phase sampling design is often used; let R be the indicator that S is measured. Let Δ be the indicator that Y is observed, that is, $\Delta = 0$ if the subject drops out before time $\tau_{\max}$ and before experiencing the event, and $\Delta = 1$ otherwise. Let $L \equiv (R(z), R(z)S(z), Y^{\tau}(z), V^{\tau}(z), \Delta(z), \Delta(z)Y(z))$ be the potential outcomes if assigned treatment z, for $z = 0, 1$, where $S(z)$ is defined if and only if $Y^{\tau}(z) = 0$, such that $S(z) =*$ if $Y^{\tau}(z) = 1$. (Note that $Y^{\tau}(z) = 1$ and $V^{\tau}(z) = 0$ each imply $R(z) = 0$.) The observed data for a subject are $O \equiv (Z, W, R, RS, Y^{\tau}, V^{\tau}, \Delta, \Delta Y)$. The CoR power calculations are based on the N vaccine recipients observed to be at-risk at τ (those with $Z(1-Y^{\tau})V^{\tau} = 1$) and test for whether $P(Y = 1|S = s_1, Z = 1, Y^{\tau} = 0)$ varies in s_1. To understand our approach, it is critical to note that the CoR power calculations do not need the potential outcomes formulation, as they are based solely on the observable random variables O. The potential outcomes are used to define biomarker-specific vaccine efficacy and hence provide a way to relate CoR effect sizes to vaccine efficacy effect sizes.

To facilitate building this relationship, we assume the vaccine has no effect on the study endpoint before the biomarker sampling time τ: $P(Y^{\tau}(1) = Y^{\tau}(0)) = 1$; this assumption will be more credible and less influential for τ near baseline. This assumption is useful by ensuring that the biomarker-specific vaccine efficacy parameters measure causal effects of vaccination and for equating the CoR parameter $P(Y = 1|S = s_1, Z = 1, Y^{\tau} = 0)$ to $P(Y = 1|S = s_1, Z = 1, Y^{\tau}(1) = Y^{\tau}(0) = 0)$, which links the CoR and VE parameter types (as described in the succeeding text). Henceforth, all unconditional and conditional probabilities of $Y(z) = 1$ tacitly condition on $Y^{\tau}(1) = Y^{\tau}(0) = 0$.

2.2. *Vaccine efficacy parameters: trichotomous biomarker*

We suppose that each of the N vaccine recipients is in one of three latent/unknown baseline subgroups X, the 'lower protected' ($X = 0$), the 'medium protected' ($X = 1$), or the 'higher protected' ($X = 2$). Define the x-specific outcome risks as

$$risk_z^{lat}(x) \equiv P(Y(z) = 1|X = x) \quad \text{for} \quad x = 0, 1, 2 \quad \text{and} \quad z = 0, 1, \tag{1}$$

such that the vaccine efficacy for latent subgroup x is $VE_x^{lat} \equiv 1 - RR_x^{lat}$ with $RR_x^{lat} \equiv risk_1^{lat}(x)/risk_0^{lat}(x)$, for $x = 0, 1, 2$.

Define $P_x^{lat} \equiv P(X = x)$ for $x = 0, 1, 2$, and define the marginal risks $risk_z \equiv P(Y(z) = 1)$ for $z = 0, 1$. Then the overall vaccine efficacy VE equals

$$VE = 1 - RR = 1 - \frac{risk_1}{risk_0} = 1 - \frac{P_0^{lat} risk_1^{lat}(0) + P_1^{lat} risk_1^{lat}(1) + P_2^{lat} risk_1^{lat}(2)}{P_0^{lat} risk_0^{lat}(0) + P_1^{lat} risk_0^{lat}(1) + P_2^{lat} risk_0^{lat}(2)}. \tag{2}$$

We also define risks and vaccine efficacies for subgroups defined by $S(1)$ or by $(X, S(1))$:

$$risk_z(s_1) \equiv P\left(Y(z) = 1|S(1) = s_1\right), \quad risk_z^{lat}(x, s_1) \equiv P\left(Y(z) = 1|X = x, S(1) = s_1\right) \tag{3}$$

for $x = 0, 1, 2$, $s_1 = 0, 1, 2$ and $z = 0, 1$, and

$$VE(s_1) \equiv 1 - RR(s_1) = 1 - risk_1(s_1)/risk_0(s_1)$$
$$VE^{lat}(x, s_1) \equiv 1 - RR^{lat}(x, s_1) = 1 - risk_1^{lat}(x, s_1)/risk_0^{lat}(x, s_1).$$

The observed biomarker response $s_1 = 0$ represents a ‘low’ response in some fashion and $s_1 = 2$ a higher response, with $s_1 = 1$ an intermediate response. For example, $s_1 = 0$ could be a negative/non-response and $s_1 = 2$ a response above a pre-specified putative correlate of protection threshold. If S were measured without error, then $X = S$ such that $VE(s_1) = VE^{lat}(x, s_1)$ and the latent variable formulation would not be needed; we use it to allow measurement error to create differences in $VE(s_1)$ versus $VE^{lat}(x, s_1)$, with greater differences for noisier biomarkers (developed next).

2.3. *Accounting for measurement error in the biomarker*

To incorporate assay noise into the power/sample size calculations, we define protection-related sensitivity/specificity and false positive/negative parameters as

$$Sens \equiv P(S(1) = 2|X = 2), \qquad Spec \equiv P(S(1) = 0|X = 0), \tag{4}$$

$$FP^0 \equiv P(S(1) = 2|X = 0), \qquad FN^2 \equiv P(S(1) = 0|X = 2), \tag{5}$$

$$FP^1 \equiv P(S(1) = 2|X = 1), \qquad FN^1 \equiv P(S(1) = 0|X = 1). \tag{6}$$

The probability an observed at-risk vaccine recipient has a low or high response, $P_0 \equiv P(S = 0|Z(1 - Y^\tau)V^\tau = 1)$ or $P_2 \equiv P(S = 2|Z(1 - Y^\tau)V^\tau = 1)$, equals

$$P_0 = Spec * P_0^{lat} + FN^1 * P_1^{lat} + FN^2 * P_2^{lat}, \tag{7}$$

$$P_2 = Sens * P_2^{lat} + FP^1 * P_1^{lat} + FP^0 * P_0^{lat}. \tag{8}$$

We consider two approaches to the trichotomous biomarker power calculations. Approach 1 takes as inputs $(Sens, Spec, FP^0, FN^2, FP^1, FN^1)$, whereas Approach 2 uses an additive measurement error model for a normally distributed continuous-readout biomarker S^* and defines the values of S by $S = 0$ if $S^* \leqslant \theta_0$, $S = 2$ if $S^* > \theta_2$, and $S = 1$ otherwise, with θ_0 and θ_2 two user-specified constants with $\theta_0 < \theta_2$. In particular, for Approach 2, we consider a normally distributed latent ‘true’ biomarker X^* and link S^* to X^* by an additive classical measurement error model

$$S^* = X^* + e, \quad X^* \sim N\left(0, \sigma_{tr}^2\right), \quad e \sim N\left(0, \sigma_e^2\right), \tag{9}$$

with X^* independent of e, implying $S^* \sim N(0, \sigma_{obs}^2)$ with $\sigma_{obs}^2 = \sigma_{tr}^2 + \sigma_e^2$. Here, $\rho \equiv 1 - \sigma_e^2/\sigma_{obs}^2$ is the fraction of the variability of S^* that is potentially biologically relevant for protection and is specified to reflect the quality of the biomarker. The ‘true’ trichotomous biomarker X is defined by two percentiles of X^* that are determined mathematically by model (9) and the two percentiles θ_0 and θ_2 (Supporting Information Appendix B). Figure 1 illustrates the set-up for Approach 2.

The aforementioned set-up handles a dichotomous biomarker as a special case, by setting $P_1^{lat} = P_1 = 0$, in which case only the *Sens* and *Spec* parameters are needed for the calculations [because $FN^2 = 1 - Sens$ and $FP^0 = 1 - Spec$; Equations (4)–(8)]. The R code handles the dichotomous biomarker as a special case.

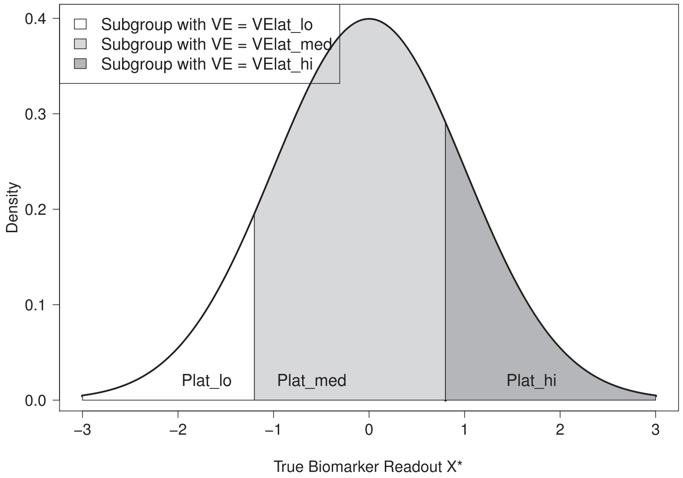

Figure 1. Division of participants into three latent subgroups with low, medium, and high levels of vaccine efficacy VE_0^{lat}, VE_1^{lat}, and VE_2^{lat}, with prevalences P_0^{lat}, P_1^{lat}, and P_2^{lat}, respectively. Under Approach 2 of Step 7 in Section 3.1 the latent normal measurement error model (9) is used.

2.4. *Vaccine efficacy parameters and model: continuous biomarker*

The formulation for a continuous biomarker is similar, where now the latent subgroups are defined by the true unobservable biomarker X^* in model (9) earlier. Now

$$VE^{lat}(x^*) \equiv 1 - risk_1^{lat}(x^*)/risk_0^{lat}(x^*), \qquad VE(s_1) \equiv 1 - risk_1(s_1)/risk_0(s_1),$$

with $risk_z^{lat}(x^*) \equiv P(Y(z) = 1|X^*(1) = x^*)$ and $risk_z(s_1) \equiv P(Y(z) = 1|S^*(1) = s_1)$ for x^* and s_1 varying over the continuous support of $X^*(1)$ and $S^*(1)$, respectively.

For the power calculations, we specify a fraction $P_{lowestVE}^{lat}$ of subjects with the lowest $X^*(1)$ values $\leqslant v$ to all have the same specified lowest level of vaccine efficacy VE_{lowest}:

$$VE_{lowest} \equiv VE^{lat}(X^*(1) \leqslant v) = 1 - risk_1^{lat}(v)/risk_0^{lat}(v). \tag{10}$$

For example, VE_{lowest} may be set to 0 and $P_{lowestVE}^{lat}$ defined as the fraction of subjects without a positive vaccine-induced immune response. The constant v is determined by $P_{lowestVE}^{lat}$, VE_{lowest}, and the measurement error model (9): $v = \sqrt{\rho}\sigma_{obs}\Phi^{-1}(P_{lowestVE}^{lat})$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cdf.

For $x^* \leqslant v$, $risk_1^{lat}(x^*)$ is modeled as a constant following (10),

$$risk_1^{lat}(x^*) = (1 - VE_{lowest})risk_0^{lat}(v) \qquad \text{for} \quad x^* \leqslant v, \tag{11}$$

and, for $x^* > v$, $risk_1^{lat}(x^*)$ is modeled via a logistic regression model

$$logit\left(risk_1^{lat}(x^*)\right) = \alpha^{lat} + \beta^{lat}x^* \qquad \text{for} \quad x^* > v. \tag{12}$$

Using model (11)–(12) that specifies a lowest value of vaccine efficacy is useful because the alternative simpler model that would specify (12) for all x would force $VE(x)$ to be negative for the lowest values of x. In many applications, this is undesirable as enhanced risk of disease caused by vaccination may be considered unlikely and the most relevant power calculations would dissallow this possibility. (Albeit the power calculator works for VE_{lowest} specified negative.)

Model (11)–(12) combined with (9) and the assumption $risk_0^{lat}(x) = risk_0$ as stated in Section 2.7 implies that

$$VE = 1 - \left[P_{lowestVE}^{lat} risk_1^{lat}(v) + \int_v^{\infty} logit^{-1}\left(\alpha^{lat} + \beta^{lat}x^*\right)\phi\left(x^*/\sqrt{\rho}\sigma_{obs}\right)dx^*\right]/risk_0, \tag{13}$$

where $\phi(\cdot)$ is the standard normal pdf. This formula will be used later for implementing the power calculations.

2.5. Correlate of risk hypotheses and estimands of interest

We address the scientific objective to assess a CoR among vaccine recipients. For trichotomous S, this entails testing the following null versus alternative hypotheses:

$$H_0 : risk_1(s_1 = 2) = risk_1(s_1 = 1) = risk_1(s_1 = 0) \quad \text{vs.} \tag{14}$$

$$H_1 : risk_1(s_1 = 2) \leqslant risk_1(s_1 = 1) \leqslant risk_1(s_1 = 0) \tag{15}$$

with '<' for at least one of the two inequalities in H_1. For continuous S^*, this tests

$$H_0 : risk_1(s_1) \text{ is constant in } s_1 \text{ vs. } H_1 : risk_1(s_1) \leqslant risk_1(s_1') \text{ for all } s_1' < s_1 \tag{16}$$

with '<' for some $s_1' < s_1$. While for data analysis, two-sided tests would typically be used, the power calculations are clearer to interpret by testing for the one-sided alternative H_1 of lower clinical risk in vaccine recipients with increasing s_1.

2.6. Methods of analysis with Bernoulli and without replacement sampling

Two main approaches to selecting the subset of subjects for whom to measure the biomarkers are

Prospective case-cohort [1]: Select a simple or stratified random sample from all randomized vaccine recipients, and augment the sample with all study endpoint cases that were not randomly sampled; and

Retrospective case-control or 2-phase sampling [4]: Conditional on final case status and possibly a discrete stratification covariate measured in all subjects, select a fixed number of vaccine recipients from each case status × covariate stratum.

Our sample size calculations consider both approaches. The first approach has advantages including that the randomly sampled subjects can be used for unbiased assessment of the distribution of the biomarker in the study population, absolute risks can be assessed in biomarker subgroups, and the association of biomarkers with multiple study endpoints can be straightforwardly assessed. The second approach does not facilitate the latter two goals, and some re-weighting is required to use the sampled subjects for unbiased assessment of the distribution of the biomarker in the study population (Supporting Information Appendix A). An advantage of the second approach is that waiting until the primary analysis is completed before selecting the controls allows accounting for the vaccine efficacy results for optimizing the biomarker sampling design. This affords opportunities to improve efficiency of the analysis [4].

2.7. Identifiability assumptions

In addition to assuming iid random variables (L_i, X_i^*, X_i) and (O_i, X_i^*, X_i) for $i = 1, \ldots, N$, we assume the standard set of assumptions that have been used in correlates of risk and protection studies: SUTVA, ignorable treatment assignment ($Z \perp L|W$), equal early clinical risk ($P(Y^\tau(1) = Y^\tau(0)) = 1$), and random censoring ($Y(z) \perp \Delta(z)$ for $z = 0, 1$). We also assume S is missing at random (MAR): R depends only on the observed data O. To the extent the investigator controls the biomarker sampling design, MAR is guaranteed to hold, although it could be in question due to happenstance missingness caused by not attending the visit at τ. Moreover, we focus on the scenario that after accounting for the latent category (and any baseline covariates W included in the CoR analysis) the measured biomarker in vaccine recipients does not affect risk, that is, $risk_1^{lat}(x^*, s_1) \equiv P(Y(1) = 1|X^*(1) = x^*, S^*(1) = s_1) = risk_1^{lat}(x^*)$ for all s_1 and x^*, and similarly for risk as a function of trichotomous X and S.

We develop the power calculations for the relatively simple scenario of homogeneous risk in the placebo group, where $risk_0^{lat}(x^*, s_1) = risk_0(s_1) = risk_0$ for all s_1 and x^* and similarly for risk as a function of trichotomous X and S. In general, $risk_0(x^*, s_1)$ and $risk_0(s_1)$ are not identifiable (because $S(1)$ is a counterfactual random variable for subjects assigned $Z = 0$), and power calculations could be conducted under many scenarios for these functions. However, the special case is very helpful for power calculations because $risk_0$ can be specified based on the observed or projected incidence in the trial. Because the CoR data analysis itself would control for known baseline prognostic factors W, the scenario in which the power calculations are accurate is $risk_0^{lat}(x^*, s_1) = risk_0(s_1) = risk_0$ after conditioning on W.

2.8. *Correlate of risk effect sizes RR_t and RR_c as a function of vaccine efficacies*

From (14), (15), and (16), analysis of the vaccine group data provides inference on the relative risks $RR_t \equiv risk_1(2)/risk_1(0)$ for a trichotomous biomarker and $RR_c \equiv risk_1(s_1)/risk_1(s_1 - 1)$ for a continuous biomarker. We refer to RR_t and RR_c as the user-specified 'CoR effect sizes' of the power calculations. From the assumptions of Section 2.7, RR_t and RR_c are identified from the observed data measured from the subset of vaccine recipients with $R = 1$, because they imply $risk_1(s_1) = P(Y = 1|Z = 1, R = 1, s_1)$ [21]. Therefore, under the assumptions the power calculations for testing H_0 can be based on the set of vaccine recipients with S (or S^*) measured at τ.

For a trichotomous biomarker, straightforward calculation shows that RR_t is linked to the latent VE parameters via the equation

$$RR_t = \frac{RR_0^{lat} * FP^0 + RR_1^{lat} * FP^1 + RR_2^{lat} * Sens}{RR_0^{lat} * Spec + RR_1^{lat} * FN^1 + RR_2^{lat} * FN^2}. \tag{17}$$

This formula makes the estimable RR_t interpretable in terms of a gradient in vaccine efficacies, where $RR_t = RR_2^{lat}/RR_0^{lat}$ for a noise-free biomarker with $1 - Sens = 1 - Spec = FP^0 = FP^1 = FN^2 = FN^1 = 0$ (illustrated in Figure 5). Otherwise, under H_1, RR_t is closer to 1.0 than RR_2^{lat}/RR_0^{lat}.

For a continuous biomarker S^* following model (9), RR_c is linked to the latent vaccine efficacy parameters via an equation that depends on s_1. Because RR_c depends on s_1, it is not particularly useful to index power calculations by RR_c. Instead, we interpret RR_c as the effect size for a noise-free biomarker ($\rho = 1$). Under the logistic model (12), RR_c is the relative risk per standard deviation increase in X^* in the region above ν, where we use the approximation of a relative risk by an odds ratio.

3. Power and sample size calculations

3.1. *Without replacement sampling*

Of the N vaccine recipients observed to be at-risk at τ, let n_{cases} ($n_{controls}$) be the number of observed cases (controls) from whom S (or S^*) is measured, where cases have $\Delta Y = 1$ and controls have $\Delta(1 - Y) = 1$. If the power calculations are done at the design stage, then N, n_{cases}, and $n_{controls}$ are projected numbers.

For a trichotomous S, the algorithm for the power calculations is as follows:

(1) Specify the overall vaccine efficacy between τ and $\tau_{\max}$, VE. For example, if the power calculations are done before the trial, then VE may be set to the protocol-specified design alternative and if afterwards to the estimated VE.

(2) Specify $risk_0$, either based on the protocol-specified placebo group endpoint rate projection or as an estimate after the trial (e.g., estimated as $n_1/(n_1 + n_2)$ where n_1 is the number of observed placebo cases between τ and τ_{max} and n_2 is the number who reached time τ_{max} free of the outcome). The estimator of $risk_0$ should be defensibly unbiased accounting for participant dropout.

(3) Select a grid of vaccine efficacies for the lower protected subgroup, VE_0^{lat}. One useful grid ranges from overall VE specified in Step 1 (the null hypothesis) to 0 (the maximal alternative hypothesis not allowing harm by vaccination).

(4) Select a grid of vaccine efficacies for the medium protected subgroup, $VE_1^{lat} \geqslant VE_0^{lat}$; one useful default choice that we use in the illustration is $VE_1^{lat} = VE$.

(5) Specify P_0^{lat} (which in many applications may be specified as the rate of 'negative' response) and specify P_2^{lat}. These values determine $P_1^{lat} = 1 - P_0^{lat} - P_2^{lat}$. Based on Equation (2) assuming $risk_0(2) = risk_0(0) = risk_0$,

$$VE = P_0^{lat} VE_0^{lat} + P_1^{lat} VE_1^{lat} + P_2^{lat} VE_2^{lat}.$$

This formula determines VE^{lat}_{hiVE} to be

$$VE^{lat}_2 = \left[VE * \left(P^{lat}_0 + P^{lat}_2\right) - P^{lat}_0 * VE^{lat}_0\right] / P^{lat}_2.$$

If one varies VE^{lat}_0 from VE to 0 as suggested in Step 3, then by the aforementioned equation VE^{lat}_2 varies from VE to $VE * (P^{lat}_0 + P^{lat}_2)/P^{lat}_2$.

(6) Specify values P_0 and P_2, which determines $P_1 = 1 - P_0 - P_2$. Thus, in total, the user specifies $(VE, risk_0, VE^{lat}_0, VE^{lat}_1, P^{lat}_0, P^{lat}_2, P_0, P_2)$. Typically, it is of interest to set $P_0 = P^{lat}_0$ and $P_2 = P^{lat}_2$ as a key scenario for computing power, as in the illustration.

(7) **Approach 1:** Following Equation (7), specify two of the three parameters $(Spec, FN^2, FN^1)$, which determines the remaining parameter. Similarly, following Equation (8), specify two of the three parameters $(Sens, FP^0, FP^1)$, which determine the remaining parameter. One choice specifies $Sens$ and $Spec$ and sets $FN^2 = FP^0 = 0$, because a biomarker of reasonable quality should have FN^2 and FP^0 close to zero.

Note that Approach 1 provides a completely general approach to studying a trichotomous or dichotomous biomarker as a CoR, without making any use of the normal measurement error model (9).

Approach 2: Specify σ^2_{obs} and ρ, which together with the values $(P_0, P_2, P^{lat}_0, P^{lat}_2)$ fixed in Step 6 determine θ_0 and θ_2 and determine $(Sens, Spec, FP^0, FN^2, FP^1, FN^1)$ (see Table I in the illustration). Typically, σ^2_{obs} is set to 1.0 as S^* can always be scaled to have unit variance.

With Approach 2, it is helpful to plot the CoR effect size RR_t versus the latent protection effect size RR^{lat}_2/RR^{lat}_0 via formula (17) for different values of ρ (illustrated in Figure 5).

(8) Simulate a large number of data sets under the aforementioned true parameter values. The procedure fixes the total number of cases n_{cases} and the total number of controls $n_{controls} = Kn_{cases}$ for an input counting number K, such that S (or S^*) are the randomly generated variables.

(9) For each simulated data set, compute the Wald test statistic for H_0 from a logistic regression model (with S the covariate of interest) that uses inverse probability weighting to account for the marker sub-sampling design. Any method for valid testing of the biomarker-outcome association using MAR case-cohort/case-control/two-phase sampling may be used in this step.

(10) Compute the power as the fraction of the simulated data sets where the Wald test statistic yields 1-sided $p \leqslant \alpha/2$ for specified α with direction favoring H_1.

(11) Given the specified VE and $risk_0$, repeat the power calculations varying ρ from 1.0 (noise-free biomarker) to a value less than one reflecting a worst case for noise level of the biomarker. Figure 4 in the next section illustrates these calculations.

(12) Repeat the power calculations for different controls : case ratios K, and, if done before the trial, possibly for different values of VE and $risk_0$.

For Step 7 Approach 2, under the assumptions of Section 2 and specified values for σ^2_{obs}, ρ, and (P^{lat}_0, P^{lat}_2), the remaining inputs $Sens$, $Spec$, FP^0, FN^2, FP^1, FN^1, (θ_0, θ_2) are determined by solving Equations (4)–(6) and (7)–(8); the R code does this using stochastic integration (Supporting Information Appendix B).

Step 8 proceeds as follows. First, for each of the N vaccine recipients, determine the numbers that are in the three latent subgroups as $N_x = P^{lat}_x * N$ rounded to the nearest integers, for $x = 0, 1, 2$. Second, determine the latent class membership of each of the n_{cases} cases by a realization of a trinomial random variable with success probabilities $(P(X = 0|Y = 1, Y^\tau = 0, Z = 1), P(X = 1|Y = 1, Y^\tau = 0, Z = 1), P(X = 2|Y = 1, Y^\tau = 0, Z = 1))$, where $P(X = x|Y = 1, Y^\tau = 0, Z = 1)$ is expressed in terms of the P^{lat}_x and $risk_1(x)$ via Bayes rule. This determines the number of cases $n_{cases}(x)$ in each category $x = 0, 1, 2$ satisfying $n_{cases} = \sum_{x=0}^2 n_{cases}(x)$. Third, within each subgroup x, simulate S_i for the entire set of N subjects as a trinomial random variable. For $x = 0$, the response probabilities are $(Spec, 1 - FP^0 - Spec, FP^0)$; for $x = 1$ the response probabilities are $(FN^1, 1 - FP^1 - FN^1, FP^1)$; and for $x = 2$, the response probabilities are $(FN^2, 1 - Sens - FN^2, Sens)$. This determines the number of controls $n_{controls}(x)$ in each category $x = 0, 1, 2$ by subtracting off $n_{cases}(x)$, satisfying the constraint $n_{controls} = \sum_{x=0}^2 n_{controls}(x)$ where $n_{controls}$ is fixed at $K * n_{cases}$. Fourth, finalize the analysis data set by specifying $R_i = 1$ or $R_i = 0$ for each of the N subjects.

For Step 9, we use the tps$(\cdot)$ function in the R package osDesign that implements the two-phase logistic regression method of Breslow and Holubkov [22], entering S as an ordered score variable with levels

Table I. Mapping of the measurement error model (9) with $\sigma^2_{obs} = 1$ indexed by ρ to $Sens$, $Spec$, FP^0, FN^2, FP^1, FN^1.

ρ	P^{lat}_0	P_0	P^{lat}_2	P_2	$Sens$	$Spec$	FP^0	FN^2	FP^1	FN^1
1	0.1	0.1	0.1	0.1	1	1	0	0	0	0
0.9	0.1	0.1	0.1	0.1	0.78	0.77	0	0	0.027	0.029
0.7	0.1	0.1	0.1	0.1	0.58	0.63	0	0	0.052	0.046
0.5	0.1	0.1	0.1	0.1	0.46	0.48	0.001	0.001	0.067	0.065
1	0.2	0.2	0.2	0.2	1	1	0	0	0	0
0.9	0.2	0.2	0.2	0.2	0.83	0.82	0	0	0.057	0.06
0.7	0.2	0.2	0.2	0.2	0.68	0.68	0.001	0.001	0.10	0.11
0.5	0.2	0.2	0.2	0.2	0.58	0.57	0.008	0.009	0.14	0.14
1	0.3	0.3	0.3	0.3	1	1	0	0	0	0
0.9	0.3	0.3	0.3	0.3	0.85	0.85	0	0	0.11	0.11
0.7	0.3	0.3	0.3	0.3	0.73	0.74	0.010	0.011	0.20	0.19
0.5	0.3	0.3	0.3	0.3	0.64	0.64	0.041	0.042	0.24	0.24
1	0.4	0.4	0.4	0.4	1	1	0	0	0	0
0.9	0.4	0.4	0.4	0.4	0.88	0.87	0.008	0.01	0.23	0.23
0.7	0.4	0.4	0.4	0.4	0.78	0.78	0.062	0.061	0.32	0.32
0.5	0.4	0.4	0.4	0.4	0.70	0.70	0.13	0.12	0.36	0.36

$S = 0, 1, 2$ and conducting a one degree of freedom Wald test. Alternatively, a generalized two degree of freedom Wald test could be used. In addition, alternative analysis methods could be used that leverage correlations between the biomarker and auxiliary covariates measured in everyone, potentially increasing power [4]. However, it will often be advantageous to base the power calculations on the simpler method both for the utility of having conservative power calculations and because the strength of correlation of the auxiliaries must be fairly high to yield a material power gain (often not available in practice).

For a continuous normally distributed biomarker S^* scaled to have mean 0 and $\sigma^2_{obs} = 1$, the same simulated data sets (using Approach 2 in Step 7) can be used for the power calculations, with process as follows.

(1) Steps 1 and 2 as for the trichotomous biomarker case.

(2) Fix $P^{lat}_{lowestVE}$, VE_{lowest}, and $\rho \in (0, 1]$. Under the logistic regression model (12) for $risk^{lat}_1(x)$, $VE^{lat}(x) = 1 - logit^{-1}(\alpha^{lat} + \beta^{lat}x)/risk_0$ for $x > \nu$ and $VE^{lat}(x) = VE_{lowest}$ for $x \leqslant \nu$. The fixed values $(VE, risk_0, P^{lat}_{lowestVE}, VE_{lowest})$ mathematically determine α^{lat} and β^{lat} by the aforementioned equation and Equation (12) (solutions in Supporting Information Appendix B).

(3) Given the specified $(VE, risk_0, P^{lat}_{lowestVE}, VE_{lowest})$, plot the VE curve $VE^{lat}(x)$ verus x given a true CoR effect size $RR_c = exp(\beta^{lat})$. $VE^{lat}(x)$ is calculated using models (11)–(12), where given fixed β^{lat}, $\alpha^{lat} = logit(risk^{lat}_1(\nu)) - \beta^{lat}\nu$. Repeat the analysis for multiple values of RR_c (see the illustration in Figure 3).

(4) Data sets are simulated by first specifying $risk^{lat}_1(x^*)$ following models (11)–(12) on a fine grid of x^* values. Second, the true latent biomarkers X^*_i for the n_{cases} cases are sampled from $P(X^* = x^* | Y = 1, Y^\tau = 0, Z = 1)$ calculated using Bayes rule. Similarly, the X^*_i values for the $n_{controls} = K * n_{cases}$ controls are sampled from $P(X^* = x^* | Y = 0, Y^\tau = 0, Z = 1)$.

(5) Steps 9–12 as mentioned earlier, except that Step 9 now uses S^* as the covariate of interest in the logistic regression model, again implemented with tps($\cdot$).

3.2. Bernoulli sampling

Under Bernoulli sampling, of the N vaccine recipients observed (or projected) to be at-risk at τ, n_{cases} ($n_{controls}$) is the expected number of observed cases (controls) from whom S and S^* are measured, that is, n_{cases} and $n_{controls}$ are random. For a trichotomous biomarker, the power analysis proceeds as described in Section 3.1, except Step 8 uses Bernoulli sampling (classic case-cohort sampling [1]). In particular, for each of the N vaccine recipients, determine the case status Y conditional on $X^* = x^*$ as a realization of a Bernoulli random variable with success probability $risk^{lat}_1(x^*)$. For a continuous biomarker, Step 8 is altered by determining the case status Y conditional on $X^* = x^*$ as a realization of a Bernoulli random variable with success probability $risk_1^{lat}(x^*)$.

4. Illustration of the power calculations

We first illustrate the correlates power calculations for the RV144 preventive HIV vaccine efficacy trial of a candidate vaccine versus placebo that was conducted in the general population in Thailand [23]. For this example, the power calculations are conducted after the trial was completed (e.g., [24, 25]). The RV144 trial randomized 8198 (8197) HIV uninfected individuals to receive vaccine (placebo) and followed them for the primary endpoint of HIV infection over 42 months. Subjects received immunizations at week 0, 4, 8, 24, and immune response biomarkers measured at τ = month 6 (week 26 visit) were assessed in vaccine recipients as CoRs of HIV infection by $\tau_{\max}$ = 42 months. Relevant for the CoR power calculations, estimated overall VE to prevent infections after τ through $\tau_{\max}$ was 0.26. Of the $N = 7703$ vaccine recipients observed to be at risk at Month 6, biomarkers were measured in the 41 subjects who were observed to subsequently experience the HIV infection endpoint, and in a frequency matched 5:1 controls : cases allocation random sample of 205 observed controls (i.e., without replacement two-phase sampling). Based on these data, several papers have reported significant continuous, trichotomous, and dichotomous CoRs in the vaccine group, with initial paper Haynes *et al.* [25]. These analyses were done using two-phase logistic regression [22] and two-phase Cox regression [26], which gave almost identical answers.

For the power analysis with a continuous biomarker following model (12), we assume $P^{lat}_{lowestVE} = 0.40$, such that the 40% of vaccine recipients with lowest X^* responses had vaccine efficacy VE_{lowest}. We varied VE_{lowest} from 0 to the overall VE estimate of 0.26. We estimated $risk_1$ as $n_1/(n_1 + n_2)$ where n_1 is the number of vaccine recipients observed to be at-risk at τ = 6 months who were diagnosed with HIV infection by the end of follow-up τ_{max} = 42 months ($n_1 = 41$) and n_2 is the number of vaccine recipients observed to be at-risk at τ who completed follow-up HIV negative ($n_2 = 7662$). Then we estimated $risk_0$ as $\widehat{risk}_1/(1 - 0.26) = 0.0072$.

Figure 2 shows the power curves for $\rho = 1, 0.9, 0.7, 0.5$. As expected, power decreases with the degree of noise. The interpretation of the plot may be aided by annotating it with results from previous efficacy trials that identified CoRs. In particular, suppose a previous trial reported an estimated $\widehat{RR}$ per sd increment in observed S^*. Under the measurement error model (9), for $\rho = 1$, this equates to $\widehat{RR}$ per sd

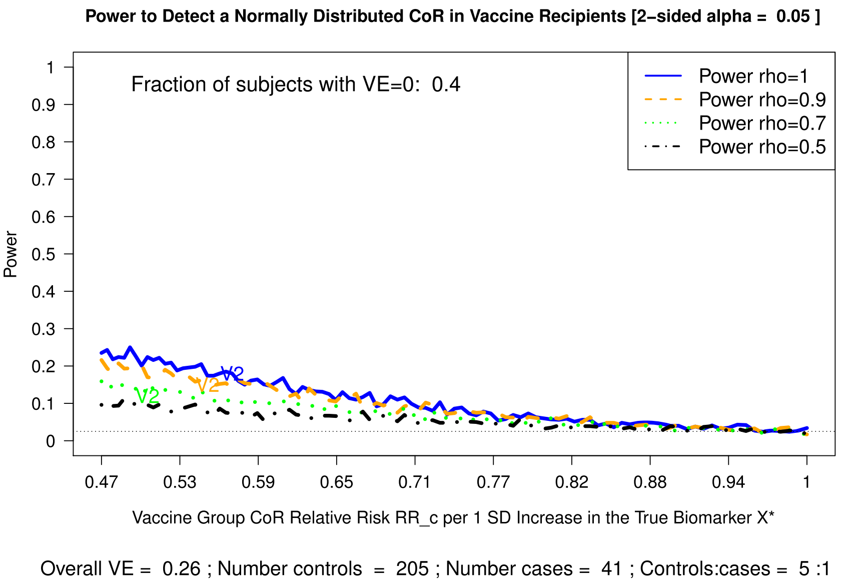

Figure 2. Power to detect a normally distributed biomarker S^* as an inverse correlate of risk (CoR) of HIV infection in vaccine recipients for the completed RV144 HIV vaccine efficacy trial with $n_{cases} = 41$ and $n_{controls} = 205$ (1-sided $\alpha = 0.05/2$-level Wald test), for the fraction ρ of the protection relevant variability of S^* ranging from 1.0 (noise-free biomarker) to 0.50 (noisy biomarker). The analysis sets $P^{lat}_{lowestVE} = 0.40$ and varies VE_{lowest} from 0 to $VE = 0.26$.

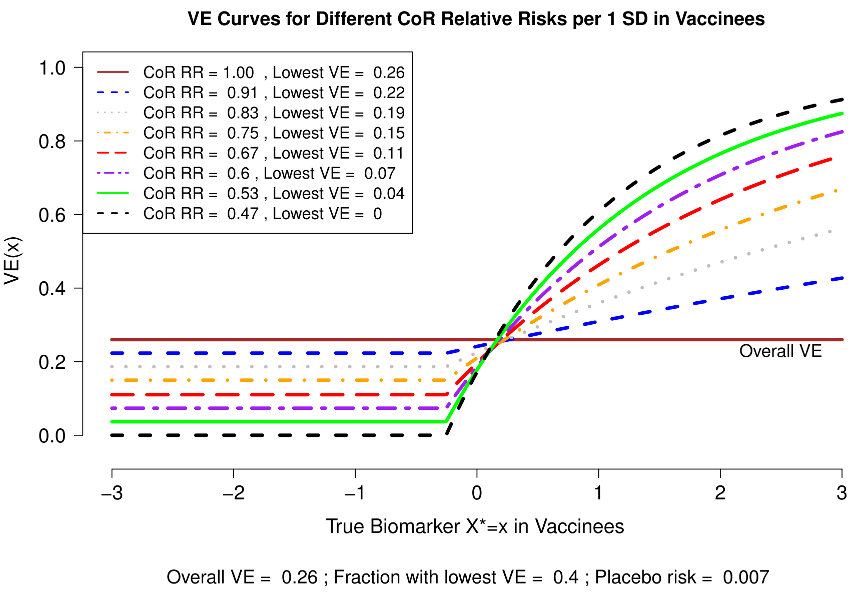

Figure 3. For the RV144 scenario with overall $VE = 0.26$, the vaccine efficacy curve $VE(x)$ for the true biomarker $X^* = x$ for six different scenarios of the true correlate of risk (CoR) relative risk effect size RR_c in the vaccine group and values of VE_{lowest} (all curves assume $\rho = 1$).

increment in X^*, and for fixed $\rho < 1$, this equates to $\widehat{RR}^{1/\sqrt{\rho}}$ per sd increment in X^*. Typically, there will be uncertainty as to the level of ρ in the previous trial, such that affixing the estimated relative risk per sd increment in X^* to each curve provides a scenario analysis of the power available to detect the previously identified CoR under a spectrum of noise levels. In Figure 2, we use the gp70-V1V2 binding antibody correlate observed in RV144 [25], which had $\widehat{RR} = 0.57$ per sd increment in S^*. If this biomarker is assumed to have no measurement error ($\rho = 1$), power is 0.19, whereas under substantial measurement error ($\rho = 0.7$), power drops to 0.13. In additional simulations, the power curves are higher if overall VE is higher (not shown).

To help interpret the power results in Figure 2, Figure 3 shows the $VE(x)$ curves for six different scenarios of the true CoR relative risk effect size RR_c ($\rho = 1$) and values of VE_{lowest} for the RV144 scenario with estimated overall VE of 0.26. The null hypothesis $RR_c = 1$ corresponds to a flat curve $VE(x) = VE$, and increasing departures from the null hypothesis H_0 correspond to increasingly variable and steep VE curves. This figure shows that for the scenario $risk_0(s_1, x) = risk_0$ and no measurement error, an association of the biomarker with infection risk in the vaccine group (a CoR) is equivalent to an association of the biomarker with VE. For interpreting Figure 2, if we focus on the $\rho = 0.9$ curve with effect size $RR_c = 0.53$ and $VE_{lowest} = 0.04$ (green solid curve), VE varies substantially in X^* but power is low, only about 0.14.

Figure 4 shows the power curves (top panels) based on the same simulated data sets following the recipe given in Section 3.1 (using Approach 2 in Step 7), for a trichotomous biomarker with $P_0 = P_2 = P_0^{lat} = P_2^{lat}$ set to 0.1, 0.2, 0.3, or 0.4 with $RR_1^{lat} = RR_{overall}$ and RR_2^{lat} tied to RR_0^{lat} through the relationship expressed in Step 5 of Section 3.1. The results show that power majorly increases with $P_0 = P_2$, which is intuitively expected given that greater sample sizes at the poles of lowest and highest VE should yield the greatest power.

To help interpret the power results of Figure 4, Figure 5 shows the relationship between the CoR effect size RR_t and the relative risk ratio RR_2^{lat}/RR_0^{lat} for the four values of ρ, with Table I showing how ρ maps to $Sens$, $Spec$, FP^0, FN^2, FP^1, FN^1 for each set of input parameters used in Figure 4. Figure 5 shows that for a noise-free biomarker with $\rho = 1$, $RR_t = RR_2^{lat}/RR_0^{lat}$ such that a CoR in the vaccine group is equivalent to the relative vaccine efficacy parameter, whereas for imperfectly measured biomarkers with $\rho < 1$, $RR_t > RR_2^{lat}/RR_0^{lat}$ such that the CoR effect size is closer to the null than the relative vaccine efficacy parameter. We illustrate a co-interpretation of Figures 4 and 5 for the $\rho = 0.9$ marker in Figure 5 and $P(S = 0) = 0.4$ (bottom-right panels). There is about 25% power to detect a CoR with effect size $RR_t = 0.60$ (Figure 4), which corresponds to 25% power to detect $RR_2^{lat}/RR_0^{lat} \approx 0.50$ (Figure 5). Supporting Information Figure 1 shows an ROC curve (sensitivity versus one minus specificity) as

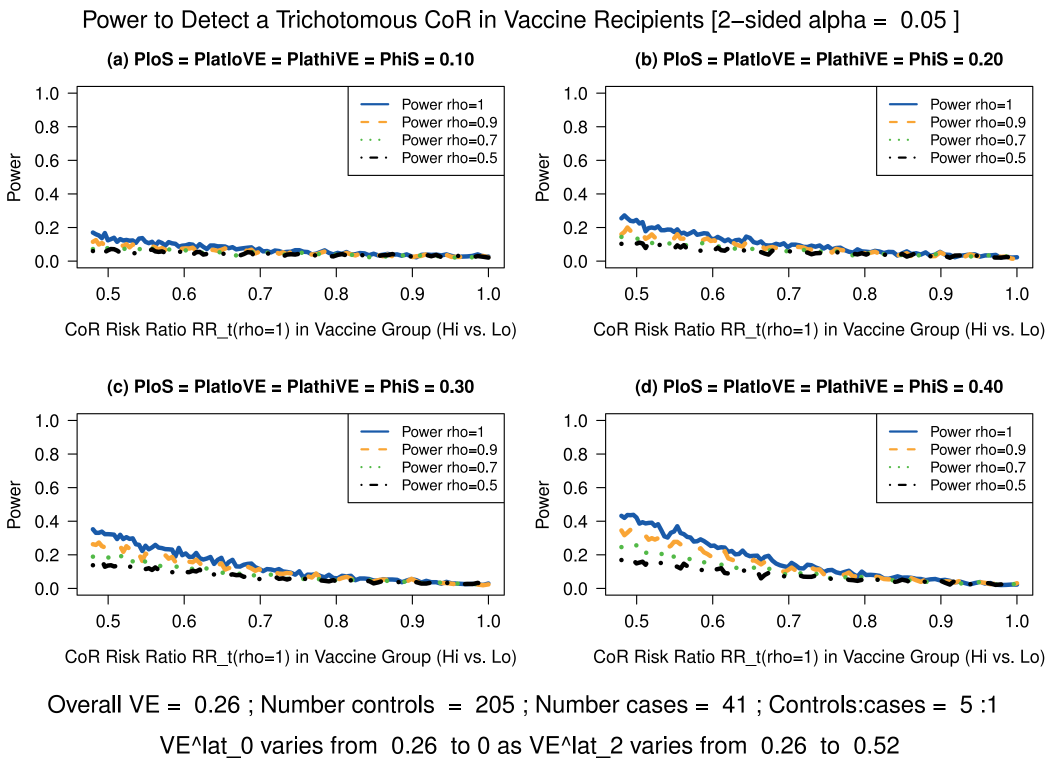

Figure 4. Vaccine group correlate of risk (CoR) power calculations for a trichotomous biomarker S for the completed RV144 HIV vaccine efficacy trial with $n_{cases} = 41$ and $n_{controls} = 205$ (1-sided $\alpha = 0.05/2$-level Wald test), for four scenarios of ρ. The x-axis is $risk_1^{lat}(2)/risk_1^{lat}(0)$, which equals RR_t for a perfect marker with $\rho = 1$.

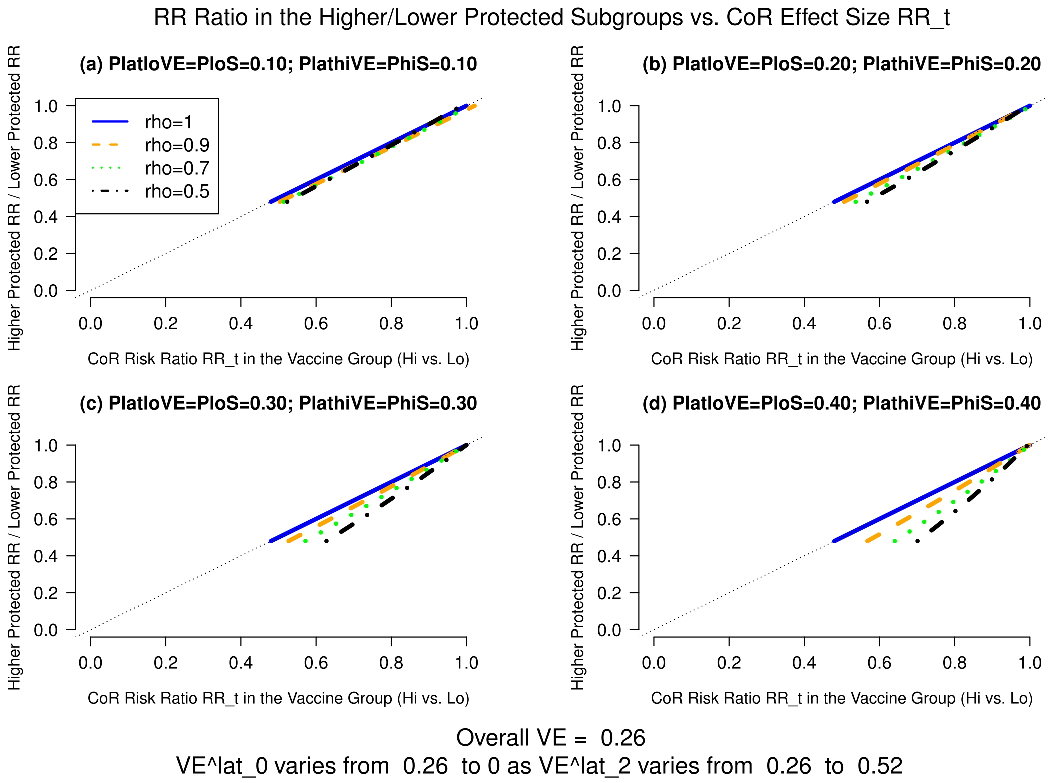

Figure 5. For the RV144 scenario with overall $VE = 0.26$, correlate of risk (CoR) effect size $RR_t = risk_1(2)/risk_1(0)$ versus the ratio RR_2^{lat}/RR_0^{lat} measuring relative vaccine efficacy for the higher protected and lower protected latent subgroups, for four scenarios of ρ.

$P_2^{lat} = P_2 = 1 - P_0 = 1 - P_0^{lat}$ ranges from 0.10 to 0.90. Our overall conclusion for this example is as follows: Because estimated overall VE was low (at 0.26), the assumption of $VE \geqslant 0$ for all biomarker subgroups constrains the possible CoR effect sizes to a limited range, hence yielding low power of the CoR analysis; in contrast, if VE were allowed to be negative for some subgroups, then power would be greater.

Our second example considers calculations being used to plan the sample size of a Phase 3 HIV vaccine efficacy trial under design by the HIV Vaccine Trials Network. This trial randomizes HIV negative individuals to vaccine or placebo in a 1:1 allocation and follows subjects for HIV infection during a $\tau_{\max} = 36$-month follow-up period. We assume 4% annual HIV incidence in the placebo group and 5% annual dropout incidence, as well as overall $VE = 0.50$. The immune response biomarkers to assess as CoRs are measured at month $\tau = 6.5$. All vaccine group subjects diagnosed with HIV between month 6.5 and 36 have biomarkers measured, as do a random sample of HIV uninfected controls with controls : cases ratio 1:1, 3:1, 5:1, or 10:1. Figure 6 shows the trichotomous biomarker power curves versus the number of infections in the vaccine group (and the total sample size observed to be at risk at τ) to detect a CoR effect size of $risk_1^{lat}(0) = 0.065$, $risk_1^{lat}(1) = 0.043$, $risk_1^{lat}(2) = 0.022$, for ρ fixed at value 0.9 that may be a realistic scenario for a biomarker assessed as a CoR. (Under the constant placebo risk assumption these calculations assume $VE_0^{lat} = 0.25$, $VE_1^{lat} = 0.50$, $VE_2^{lat} = 0.75$.) The calculations are for the scenarios $P_0 = P_0^{lat} = P_2^{lat} = P_2$ ranging from 0.10 to 0.50. The results show that power sharply increases with the prevalences $P_0^{lat} = P_2^{lat}$ and increases with the controls : cases ratio, with only incremental gain moving from 5:1 to 10:1. Based on this analysis, to achieve 90% power to detect a CoR with $P_0 = P_2 = 0.30$, one choice would be the 5:1 allocation design, requiring 2800 total vaccine recipients observed to be at-risk at 6.5 months.

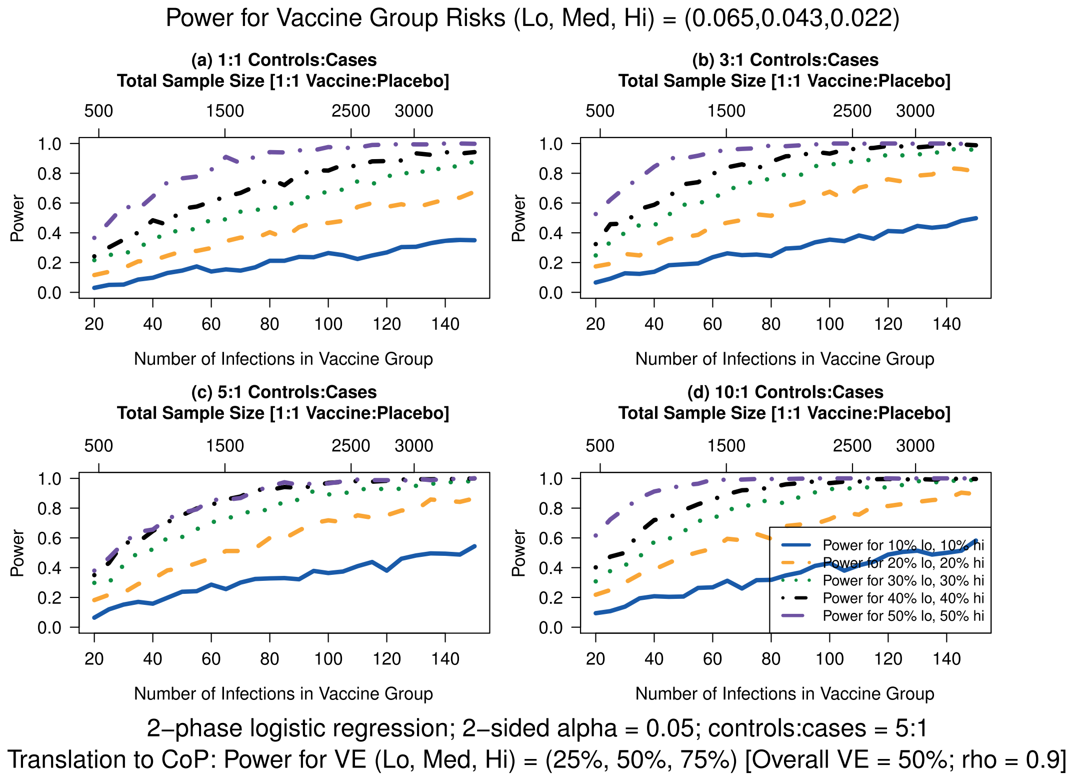

Figure 6. Vaccine group correlate of risk power curves versus total sample size (number of participants observed to be at risk at time τ) for a trichotomous biomarker S measured at time τ to plan a two-arm HIV vaccine efficacy trial with equal allocation randomization to vaccine versus placebo, 4% annual placebo incidence, 5% annual dropout incidence, overall $VE = 0.50$, $\rho = 0.9$, and CoR effect size $risk_1^{lat}(0) = 0.069$, $risk_1^{lat}(1) = 0.043$, and $risk_1^{lat}(2) = 0.013$. Each panel shows power for $P_0 = P_0^{lat} = P_2^{lat} = P_2$ ranging from 0.1 to 0.5, for controls : cases allocation ratios (a) 1:1, (b) 3:1, (c) 5:1, and (d) 10:1. Assuming $risk_0(x, s_1) = risk_0$, the corresponding vaccine efficacy effect size is $VE_0^{lat} = 0.25$, $VE_1^{lat} = 0.50$, $VE_2^{lat} = 0.75$. The total sample size is the number of participants observed to be at risk at time τ (vaccine + placebo).

5. Discussion

We developed an approach to power and sample size calculations for a typical ‘CoR’ data analysis in a randomized controlled clinical efficacy trial for testing an association of an observed biomarker measured in a sub-sample (via a case-cohort, case-control, or two-phase sampling design) of the active treatment group with a clinical endpoint. The contribution of this work is to integrate into the calculations two issues – the level of treatment efficacy across biomarker subgroups and the fraction ρ of the variability of the biomarker that is potentially biologically relevant for protection ($\rho \equiv 1 - \sigma_e^2/\sigma_{obs}^2$). The first issue is important because, if ignored, a statistician may design the sample size of a CoR study not realizing the tacit assumptions being made about treatment efficacy. A particularly egregious mistake would be powering a study to detect a CoR with no recognition that achieving the desired power requires that treatment efficacy be negative for some biomarker subgroups, rendering the CoR study underpowered if treatment efficacy is never negative. Our approach provides a way to explicitly explore the relationship of the CoR effect size with treatment efficacy, including a way to specify the lowest treatment efficacy at a fixed value such as zero. The second issue is important because the degree of measurement error ρ heavily influences power [14–16], such that accounting for ρ is needed for accurate power calculations, and may be useful for screening out biomarkers for which the CoR study would be underpowered given an unacceptably low value of ρ.

For the continuous biomarker calculations and for the Approach 2 trichotomous biomarker calculations, we have assumed a classical additive normal measurement error model for the observed continuous biomarker S^*, the veracity of which should be tested. In general, in the planning of biomarker CoR studies, it is important to conduct biomarker assay laboratory validation studies to estimate ρ; we discuss approaches to this in Web Appendix C.

Our power calculator applies for a univariate biomarker, yet studying the association of multiple biomarkers with outcome is an important application. The calculator for a trichotomous biomarker may be useful for trials that collect possibly high-dimensional multivariate biomarkers and for which unsupervised clustering based on the biomarkers yields a cluster of ‘putatively not protected’ subjects and a cluster of ‘putatively protected’ subjects. In this scenario, the power calculator may be applied with all other subjects constituting the third cluster. In addition, the calculator for a normally distributed biomarker may be used for studying power to detect a linear combination of multiple biomarkers as a CoR.

Our CoR power and sample size calculations are for the scenario that the biomarker is not associated with the clinical endpoint in the control group after accounting for baseline covariates W that would be controlled for in the CoR data analysis. This assumption is not needed for the CoR calculations because they use data from the active treatment group only. However, this assumption is used as a way to interpret the CoR power calculations in terms of biomarker-specific treatment efficacy, providing a mapping from the CoR calculations (in terms of risk gradients in the active treatment group) to gradients in treatment efficacy. Additional calculations may be conducted under alternative scenarios, where the approach here could be extended to allow functions $risk_0(x, s_1)$ other than $risk_0(x, s_1) = risk_0$. While the main application of the methods is more interpretable and accurate CoR power and sample size calculations, a second application is power and sample size calculations for assessing modification of treatment efficacy by the biomarker, that is, assessing the vaccine efficacy curve $VE(s_1)$ directly, which is conducted in the principal stratification framework [18,27]. Supporting Information Figures S1 and S3 show such power curves for our two illustrative examples.

Acknowledgements

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Numbers R37AI054165 and UM1AI068635. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank the participants, investigators, and sponsors of the RV144 trial, including the US Military HIV Research Program (MHRP); US Army Medical Research and Materiel Command; NIAID; US and Thai Components, Armed Forces Research Institute of Medical Science; Ministry of Public Health, Thailand; Mahidol University; SanofiPasteur; and Global Solutions for Infectious Diseases.

References

1. Prentice R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1–11.
2. Barlow W. Robust variance estimation for the case-cohort design. *Biometrics* 1994; **50**:1064–1072.

3. Kulich M, Lin D. Improving efficiency of relative-risk estimation in case cohort studies. *Journal of the American Statistical Association* 2004; **99**:832–844.
4. Breslow N, Lumley T, Ballantyne C, Chambless L, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* 2009; **169**:1398–1405.
5. Cai J, Zeng D. Sample size/power calculation for case-cohort studies. *Biometrics* 2004; **60**:1015–1024.
6. Dupont WD, Plummer Jr. Power and sample size calculations: a review and computer program. *Controlled Clinical Trials* 1990; **11**:116–128.
7. García-Closas M, Lubin JH. Power and sample size calculations in case control studies of gene-environment interactions: comments on different approaches. *American Journal of Epidemiology* 1999; **149**:689–692.
8. Haneuse S, Saegusa T, Lumley T. osDesign: an R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of Statistical Software* 2011; **43**:11. pii: v43/i11/paper.
9. Gilbert PB, Grove D, Gabriel E, Huang Y, Gray G, Hammer S, Buchbinder S, Kublin J, Corey L, Self SG. A sequential Phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple HIV vaccine regimens. *Statistical Communications in Infectious Diseases* 2011; **3**(1):Article 4.
10. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics* 1999; **55**:1193–1201.
11. Tosteson TD, Buzas JS, Demidenko E, Karagas M. Power and sample size calculations for generalized regression models and covariate measurement error. *Statistics in Medicine* 2003; **22**:1069–1082.
12. Chen LM, Ibrahim JG, Chu H. Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine* 2011; **30**:2295–2309.
13. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**(260):663–685.
14. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statistics in Medicine* 1989; **8**:1151–1163.
15. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within person measure error. *Statistics in Medicine* 1989; **8**:1051–1070.
16. McKeown-Eyssen GE, Tibshirani R. Implications of measurement error in exposure for the sample sizes of case-control studies. *American Journal of Epidemiology* 1994; **139**(4):415–421.
17. Halloran ME, Longini IM, Struchiner CJ. *Design and Analysis of Vaccine Studies*. Springer: New York, 2010.
18. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics* 2008; **64**:1146–1154.
19. Qin L, Gilbert PB, Corey L, McElrath MJ, Self SG. A framework for assessing an immunological correlate of protection in vaccine trials. *The Journal of Infectious Diseases* 2007; **196**:1304–1312.
20. Plotkin S, Gilbert PB. Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Diseases* 2012; **54**:1615–1617.
21. Gabriel E, Gilbert PB. Evaluating principle surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics* 2014; **15**:251–265.
22. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1997; **59**:447–461.
23. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S. Kim JH for the MOPH-TAVEG Investigators. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine* 2009; **361**:2209–2220.
24. Rolland M, Gilbert PB. Evaluating immune correlates in HIV Type 1 vaccine efficacy trials: what RV144 may provide. *AIDS Research and Human Retroviruses* 2012; **28**:400–404.
25. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, Evans DT, Montefiori DC, Karnasuta C, Sutthent R, Liao HX, DeVico AL, Lewis GK, Williams C, Pinter A, Fong Y, Janes H, deCamp A, Huang Y, Rao M, Billings E, Karasavvas N, Robb ML, Ngauy V, de Souza MS, Paris R, Ferrari G, Bailer RT, Soderberg KA, Andrews C, Berman PW, Frahm N, De Rosa SC, Alpert MD, Yates NL, Shen X, Koup RA, Pitisuttithum P, Kaewkungwal J, Nitayaphan S, Rerks-Ngarm S, Michael NL, Kim JH. Immune correlates analysis of the ALVAC-AIDSVAX HIV-1 vaccine efficacy trial. *New England Journal of Medicine* 2012; **366**:1275–1286.
26. Borgan L, Langholz B, Samuelson S, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58.
27. Frangakis C, Rubin D. Principal stratification in causal inference. *Biometrics* 2002; **58**:21–29.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.