Statistical Communications in Infectious Diseases

Volume 2, *Issue* 1 2010 *Article* 1

Increasing the Efficiency of Prevention Trials by Incorporating Baseline Covariates

Min Zhang, *University of Michigan*
Peter B. Gilbert, *Fred Hutchinson Cancer Research Center & University of Washington*

Recommended Citation:
Zhang, Min and Gilbert, Peter B. (2010) "Increasing the Efficiency of Prevention Trials by Incorporating Baseline Covariates," *Statistical Communications in Infectious Diseases*: Vol. 2: Iss. 1, Article 1.
DOI: 10.2202/1948-4690.1002

Increasing the Efficiency of Prevention Trials by Incorporating Baseline Covariates

Min Zhang and Peter B. Gilbert

Abstract

Most randomized efficacy trials of interventions to prevent HIV or other infectious diseases have assessed intervention efficacy by a method that either does not incorporate baseline covariates, or that incorporates them in a non-robust or inefficient way. Yet, it has long been known that randomized treatment effects can be assessed with greater efficiency by incorporating baseline covariates that predict the response variable. Tsiatis et al. (2007) and Zhang et al. (2008) advocated a semiparametric efficient approach, based on the theory of Robins et al. (1994), for consistently estimating randomized treatment effects that optimally incorporates predictive baseline covariates, without any parametric assumptions. They stressed the objectivity of the approach, which is achieved by separating the modeling of baseline predictors from the estimation of the treatment effect. While their work adequately justifies implementation of the method for large Phase 3 trials (because its optimality is in terms of asymptotic properties), its performance for intermediate-sized screening Phase 2b efficacy trials, which are increasing in frequency, is unknown. Furthermore, the past work did not consider a right-censored time-to-event endpoint, which is the usual primary endpoint for a prevention trial. For Phase 2b HIV vaccine efficacy trials, we study finite-sample performance of Zhang et al.'s (2008) method for a dichotomous endpoint, and develop and study an adaptation of this method to a discrete right-censored time-to-event endpoint. We show that, given the predictive capacity of baseline covariates collected in real HIV prevention trials, the methods achieve 5-15% gains in efficiency compared to methods in current use. We apply the methods to the first HIV vaccine efficacy trial. This work supports implementation of the discrete failure time method for prevention trials.

KEYWORDS: auxiliary, covariate adjustment, intermediate-sized phase 2b efficacy trial, semiparametric efficiency

Author Notes: This work was supported by NIH grant 2 R01 AI54165-04.

1 Introduction

1.1. Motivating application: Background on Phase 2b and 3 HIV vaccine efficacy trials. Initially, in the 1980's and early 1990's, the HIV vaccine field focused on the development of candidate vaccines designed to prevent HIV infection by stimulating anti-HIV neutralizing antibodies. Two Phase 3 efficacy trials of such a candidate have been conducted, of VaxGen's recombinant envelope glycoprotein vaccine (Flynn et al., 2005; Pitisuttithum et al., 2006) (Table 1). The primary objective of these trials was to estimate the vaccine efficacy (VE), defined as one minus the hazard ratio (vaccine group/placebo group) of HIV infection diagnosis in study volunteers who were initially HIV negative.

Given that Phase 3 HIV vaccine trials are very expensive, and that initial candidate HIV vaccines have a low probability of working, in the nineties the vaccine field conceived an alternative streamlined type of efficacy trial, named a Phase 2b intermediate-sized efficacy trial (Rida et al., 1997). Phase 2b trials are generally designed to accrue between a quarter and one-third the number of infection endpoints that would be assessed in a Phase 3 trial; a prototype design has 100 total endpoints (Fleming and Richardson, 2004). Gilbert (2009) described a decision theoretic framework for comparing the expected utility of choosing an initial efficacy trial design as a Phase 2b or Phase 3 trial. The decision analysis suggests that for candidate vaccines with low pre-test plausibility of efficacy, the Phase 2b trial has greater expected utility than the Phase 3 trial under a broad range of utilities. Consistent with these results, the HIV vaccine field now focuses on Phase 2b trials, of which three have been conducted (Table 1). Two of these have final reported results, the "Step trial" (Buchbinder et al., 2008) and the RV 144 trial (Rerks-Ngarm et al., 2009), which accrued 83 and 125 HIV infections by the time of the final analysis, respectively.

Phase 2b designs are also being increasingly used for HIV prevention trials of non-vaccine interventions. For example Fleming and Richardson (2004) described implementation of a Phase 2b design for a microbicide prevention trial, which evaluated two microbicide groups versus two control groups (one blinded and one unblinded). This trial was designed to accrue approximately 100 total HIV infections for comparing each active group versus each control group. The methods studied here apply generally for double-blind, randomized Phase 2b trials with a primary endpoint that is either dichotomous or discrete time-to-event.

1.2 Improving efficiency of Phase 2b and 3 HIV prevention trials. In both Phase 2b and 3 HIV vaccine efficacy trials, the standard statistical approach for evaluating VE is to estimate a ratio (vaccine/placebo) of instantaneous or cumulative incidence rates. All of the trials described in Table 1 (with sufficient endpoint data) used a discrete-time and/or continuous-time Cox proportional hazards model to estimate VE with one minus the exponent of the maximum partial likelihood estimator of the regression parameter for the vaccination assignment. This estimator is globally semiparametric efficient for the Cox model among regular estimators that do not account for participant covariates other than vaccination assignment.

It has long been recognized that methods that account for baseline covariates predictive of the response variable can improve efficiency (e.g., Egger et al., 1985). Such methods have rarely been used for the primary analysis of a clinical efficacy trial, however, as they require modeling of the relationship between the covariates and the response variable, and the estimation of the treatment effect depends on the choice of model. This opens the analysis to an unscrupulous approach wherein the investigator selects the model such that the treatment effect 'looks good' (Pocock, 2002; Tsiatis et al., 2007), and, moreover, for many available methods the validity of estimation depends on uncertain parametric assumptions. For Phase 2b or 3 trials that may influence licensure decisions, regulatory agencies have rejected such analysis approaches that are subjective or sensitive to parametric assumptions, strongly favoring objective approaches that minimize assumptions. To overcome these barriers, Tsiatis et al. (2007) and Zhang, Tsiatis, and Davidian (2008) (henceforth ZTD) showed how the semiparametric efficient approach of Robins, Rotnitzky, and Zhao (1994) can be applied in an objective and robust manner, in which the two steps of building predictive models and solving the estimating equation for the treatment effect are completely separate. Their separation allows an objective approach either through using independent statisticians for the two steps, or through using an automated and pre-specified approach to model selection in the first step (Tsiatis et al., 2007). Moreover, the method is valid without additional assumptions beyond those made by standard analysis approaches that have long been accepted by regulatory agencies. Instead of directly solving estimating equations, Moore and van der Laan (2009) and van der Laan and Rubin (2006) studied targeted maximum likelihood method (tMLE) as an alternative approach. The tMLE methodology iteratively updates an initial density estimator with the goal of achieving bias reduction for the parameter of interest and is closely related to estimating equation based methods. In this paper, we focus on the approach that directly solves estimating equations, as in Tsiatis et al. (2007) and ZTD.

Table 1: *History of HIV Vaccine Efficacy Trials.*

Efficacy Trial	Phase	Time Period	Vaccine	Study Population	Primary Endpoint(s)	Sample Size	No. Events	Outcome
VaxGen 004	3	1998-2003	gp120 protein (antibodies)	North America Men sex w/ men (MSM)	Infection	5403 2:1 V:P	**368** 241:127	$\widehat{RR} = .94$ 95% CI .76–1.17 $p = .59$
VaxGen 003	3	1999-2003	gp120 protein	Thailand IDU	Infection	2527 1:1 V:P	**211** 106:105	$\widehat{RR} = 1.0$ 95% CI .76–1.31 $p = .99$
RV 144	2b	2004-2009	ALVAC prime: gp120 boost	Thailand General Pop^{n}	Infection	16,395 1:1 V:P	**125** 51:74 $p = .04$	$\widehat{RR} = 0.69$ 95% CI .49–.99
Step 502	2b	2004-2008	Ad5 vector	Americas MSM + Women	Infection; Viral Load	3000 1:1 V:P	**83** 49:33	$\widehat{RR} = 1.5$ 95% CI .95–2.41 $p = .07$
Step 503	2b	2006-2008	Ad5 vector	South Africa Heterosexual Men + Women	Infection; Viral Load	3000 1:1 V:P	**11**	Trial unblinded after public release of Step 502 results

Through simulation studies Tsiatis et al. (2007) and ZTD showed that, for a quantitative endpoint and for a dichotomous endpoint, respectively, the semiparametric efficient approach performs well in practice for large clinical trials. ZTD's results for a dichotomous endpoint may justify its current use for Phase 3 vaccine trials, because the "asymptotics kick-in" for Phase 3 trials, and because it is reasonable to evaluate VE using the dichotomous endpoint of HIV infection. The latter point follows because HIV infection is a rare event, implying that minimal efficiency is lost by moving from a time-to-event endpoint to a dichotomous endpoint (Cuzick et al., 1982).

However, to justify the use of the semiparametric efficient method for a dichotomous endpoint, more work is needed on two fronts. First, the method's use in Phase 2b trials is not yet justified, because the efficiency of the method is defined in terms of large sample theory. ZTD considered a non rare-event setting with approximately 275–330 events, whereas here we consider the rare-event setting with far fewer events (approximately 60–120). Moreover, in small samples the 'sandwich method' for estimating the variance of the vaccine effect can be biased. While Tsiatis et al. (2007) proposed a small-sample variance correction, its performance has not been fully vetted. Second, adaptations of the method for a right-censored time-to-event endpoint, for either Phase 2b or 3 trials, have not been fully developed or studied. While it may be acceptable to use a dichotomous endpoint in a prevention trial with a rare event, use of a time-to-event endpoint has advantages even in the rare event setting. In particular, time-to-event methods can accommodate the heterogeneity in at-risk periods of subjects arising because of staggered entry and study drop-out, and can better accommodate sequential monitoring plans. Moreover, the hazard-based estimand used with time-to-event methods may be of greater scientific interest than a final value-based estimand.

Moore and van der Laan (2009b) studied the tMLE approach to estimate the survival curve at a fixed time point. Lu and Tsiatis (2008) studied an estimating equation based method for Cox's proportional hazards model, where they used an additional assumption that, conditional on the treatment assignment, the censoring time is independent of the failure time and of baseline and time-dependent covariates. Based on this assumption, they derived estimators that incorporate both baseline and time-dependent covariates to improve efficiency. They did not focus on the objectivity of the method, and the method does not fit regression models separately for each treatment group. In this article we will develop a semiparametric method for a discrete time-to-event model that does not make this additional assumption.

The purposes of this article are to evaluate the finite-sample performance of the semiparametric efficient method for a dichotomous endpoint, and to develop and evaluate the finite-sample performance of a newly proposed semiparametric method for a discrete time-to-event endpoint, all for the setting of a Phase 2b prevention trial with a rare event. The goal is to inform the questions of if and how the primary analysis of a Phase 2b trial should take into account baseline covariates. Accounting for covariates holds potential to improve efficiency of HIV prevention trials in practice because covariates are readily available that predict the rate of HIV infection (e.g., demographic data, self-report risk behavior data, status of infection with sexually transmitted diseases). More generally, the proposed methods may provide valuable efficiency gains in Phase 2b or 3 trials of other types of interventions.

This article is organized as follows. Section 2 describes ZTD's semiparametric efficient method for a dichotomous endpoint. Section 3 proposes a new, related semiparametric method for a right-censored discrete time-to-event endpoint. Section 4 evaluates the methods in simulations designed to mimic a prototype Phase 2b vaccine efficacy trial, showing how much efficiency can be gained compared to standard methods in current use that ignore baseline predictors of the study endpoint. Section 5 applies the method to the first HIV vaccine efficacy trial, and Section 6 concludes with discussion.

2 Efficient Method for a Dichotomous Outcome

Consider a prevention trial that enrolls n initially HIV uninfected subjects, who are randomly assigned to the vaccine group ($Z = 1$) or the placebo group ($Z = 0$) with randomization probability $P(Z = 1) = \pi$. Let n_0 and n_1 be the number of subjects assigned to the placebo and vaccine groups. With a dichotomous endpoint Y (e.g., diagnosis of HIV infection during a fixed follow-up period), we define VE as one minus the odds ratio (OR) of $Y = 1$ for the vaccine versus placebo group. Let $X(p \times 1)$ be a vector of baseline covariates. Because of randomization, treatment assignment Z is independent of X, expressed as $Z \perp\!\!\!\perp X$, and this independence is the key to the development of a more efficient method using baseline covariates. The observed data are summarized as (Y_i, Z_i, X_i), $i = 1, \ldots, n$, which are supposed independent and identically distributed.

The parameter of interest VE, or equivalently the log odds ratio, can be defined through a logistic regression model that includes only the intercept and treatment assignment Z, i.e.,

$$\text{logit}\{E(Y|Z)\} = \beta_0 + \beta_1 Z, \quad \beta = (\beta_0, \beta_1)^T, \quad Y|Z \sim \text{Bernoulli}, \qquad (1)$$

where $\text{logit}(u) = \log\{u/(1-u)\}$; β_0 and β_1 are, respectively, the log odds of infection for the placebo group and the log odds ratio for the vaccine group relative to the placebo group (with $VE \equiv 1 - \exp(\beta_1)$). In this model the treatment effect β_1 is defined unconditionally on covariates. In addition, this saturated model imposes no restrictions on the conditional mean of Y given Z.

2.1. Standard maximum likelihood estimator (MLE): a function of (Y_i, Z_i), $i = 1, \ldots, n$. Standard maximum likelihood inference for model (1) is based on data on (Y, Z) only, where the MLE is obtained by solving the estimating equation

$$\sum_{i=1}^{n} m(Y_i, Z_i; \beta) = \sum_{i=1}^{n} (1 - Z_i, Z_i)^T \{Y_i - \text{expit}(\beta_0 + \beta_1 Z_i)\} = 0, \tag{2}$$

where $\text{expit}(u) = \exp(u)/\{1 + \exp(u)\}$. We refer to the summand in (2) as the estimating function for subject i, and denote it as $m(Y, Z; \beta)$, suppressing the subscript i. This estimating function is unbiased in the sense that $E_\beta\{m(Y, Z; \beta)\} = 0$. Written explicitly, the MLE for β is

$$\widehat{\beta}_{mle} = \begin{pmatrix} \text{logit}(\bar{Y}_0) \\ \text{logit}(\bar{Y}_1) - \text{logit}(\bar{Y}_0) \end{pmatrix}, \tag{3}$$

where $\bar{Y}_0$ and $\bar{Y}_1$ are the sample average of responses for the placebo and vaccine groups. Of all regular and asymptotically linear (RAL) estimators of β that are functions of (Y_i, Z_i), $i = 1, \ldots, n$, the MLE $\widehat{\beta}_{mle}$ is the most efficient.

Because $\widehat{\beta}_{mle}$ is a function of the (Y_i, Z_i) only, however, it is not necessarily most efficient among all estimators that are functions of the (Y_i, Z_i, X_i). Within the framework of semiparametric theory, ZTD derived the class of all RAL estimators for the treatment effect in a randomized trial based on (Y_i, Z_i, X_i), $i = 1, \ldots, n$; we refer to (Y_i, Z_i, X_i) as the full data and (Y_i, Z_i) as the reduced data. In this development, no additional assumptions are made about the marginal distribution of X or the joint distribution of (Y, Z, X) except those already made for the original model (1); instead, all aspects of the joint distribution of the full data that are not dictated by the design of the experiment or specified by the original model are viewed as nuisance parameters. Accordingly, the ZTD semiparametric method may be suitable for the primary analysis of a clinical trial, for which it is often desirable to minimize assumptions.

2.2. Augmented estimators: functions of (Y_i, Z_i, X_i), $i = 1, \ldots, n$. Applying the results of ZTD to our problem, all unbiased estimating functions for β

based on the full data (Y, Z, X) are of the form

$$m^*(Y, Z, X; \beta) = m(Y, Z; \beta) - (Z - \pi)a(X), \tag{4}$$

where $a(X)$ is an arbitrary 2-dimensional function of X. Moreover, $m(Y, Z; \beta)$, defined in (2), is the unbiased estimating function based on the reduced data. Because $Z \perp\!\!\!\perp X$, the second term in $m^*(Y, Z, X; \beta)$, $(Z-\pi)a(X)$, obviously has expectation zero for any function $a(X)$, and therefore $m^*(Y, Z, X; \beta)$ is again an unbiased estimating function. Under regularity conditions, an unbiased estimating function can be used to obtain a consistent and asymptotically normal estimator of β (e.g., Carroll et al., 2006, Section A.6), defined as the solution to the corresponding estimating equation

$$\sum_{i=1}^{n} m^*(Y_i, Z_i, X_i; \beta) = \sum_{i=1}^{n} m(Y_i, Z_i; \beta) - (Z_i - \pi)a(X_i) = 0, \tag{5}$$

which we refer to as an augmented estimating equation. The class of estimators indexed by $a(X)$ include the usual MLE $\widehat{\beta}_{mle}$ as a special case, achieved by setting $a(X) \equiv 0$. Although any $a(X)$ will lead to a consistent and asymptotically normal estimator for β, judicious choices of $a(X)$ will lead to more efficient estimators.

ZTD identified that the most efficient estimator in the class (5) is achieved with $a(X) \equiv E\{m(Y, Z)|Z = 1, X\} - E\{m(Y, Z)|Z = 0, X\}$, and therefore the optimal estimator for β can be obtained by solving the estimating equation

$$\begin{aligned}&\sum_{i=1}^{n} m^*_{opt}(Y_i, Z_i, X_i; \beta)\\ &= \sum_{i=1}^{n} \begin{pmatrix} (1 - Z_i)Y_i + (Z_i - \pi)E(Y_i|Z_i = 0, X_i) - (1 - \pi)\text{expit}(\beta_0) \\ Z_iY_i - (Z_i - \pi)E(Y_i|Z_i = 1, X_i) - \pi\text{expit}(\beta_0 + \beta_1) = 0 \end{pmatrix} = 0.\end{aligned} \tag{6}$$

In practice, the two conditional expectations, $E(Y|Z = 1, X)$ and $E(Y|Z = 0, X)$, are unknown and must be modeled. Fortunately, many modeling techniques are available, and a practical estimation strategy builds regression models for $E(Y|Z = 1, X) = q_1(X; \eta)$ and $E(Y|Z = 0, X) = q_0(X; \zeta)$ separately based on data from group 1 and 0, respectively, and then, using each fitted model, predicts Y for all subjects in the trial. Given the resulting predicted values $q_1(X_i; \widehat{\eta})$ and $q_0(X_i; \widehat{\zeta})$, $i = 1, \ldots, n$, the final estimator of β is obtained by replacing $E(Y_i|Z_i = 1, X_i)$ and $E(Y_i|Z_i = 0, X_i)$ in (6) with $q_1(X_i; \widehat{\eta})$ and

$q_0(X_i;\widehat{\zeta})$, and by replacing π in (6) with the sample proportion $\widehat{\pi} = n_1/n$. The resulting estimator has the closed form expression

$$\widehat{\beta}_{aug} = \left(\begin{array}{c} \widehat{\beta}_{aug0} \\ \widehat{\beta}_{aug1} \end{array} \right) = \left(\begin{array}{c} \text{logit}\{\bar{Y}_0 + n_0^{-1} \sum_{i=1}^{n} (Z_i - \widehat{\pi}) q_0(X_i;\widehat{\zeta})\} \\ \text{logit}\{\bar{Y}_1 - n_1^{-1} \sum_{i=1}^{n} (Z_i - \widehat{\pi}) q_1(X_i;\widehat{\eta})\} - \widehat{\beta}_{aug0} \end{array} \right), \tag{7}$$

where $\bar{Y}_0$, $\bar{Y}_1$ are defined previously. As the specified regression models may not be the true models, this estimator is not necessarily the optimal one, and we refer to it as an augmented estimator, denoted by $\widehat{\beta}_{aug}$. We denote the corresponding estimating function as $m^*_{aug}(Y, Z, X; \beta)$, which is the summand in (6) with $E(Y_i|Z_i = 0, X_i)$ and $E(Y_i|Z_i = 1, X_i)$ replaced with $q_0(X_i, \zeta)$ and $q_1(X_i, \eta)$.

This augmented estimator has a robustness property: if either or both of the specified regression models $q_1(X;\eta)$ and/or $q_0(X;\zeta)$ are false functional forms for the conditional expectations, the resulting estimator is still a member of the class of estimators (5), and is therefore consistent and asymptotically normal. According to ZTD, if we specify linear regression models and fit them using ordinary least squares (OLS) estimators, then it is guaranteed that the augmented estimator will be at least as efficient asymptotically as the standard estimator that ignores baseline covariate data; see also Leon et al. (2003, Section 4). If other estimation methods (e.g., iteratively reweighted least squares (IRWLS) estimators for logistic regression models) are used for the augmentation terms $q_1(X;\eta)$ and $q_0(X;\zeta)$, then a slightly modified estimating equation approach can be used, and again the resulting estimator is guaranteed to be at least as efficient asymptotically as the standard estimator (van der Laan and Robins, 2003, Chapter 2, Section 2.5). This approach entails an additional step of fitting linear regression models by OLS and we do not pursue it further in this paper.

Because the augmentation terms $q_1(X;\eta)$ and $q_0(X;\zeta)$ are conditional expectations of a dichotomous variable, a natural modeling approach is logistic regression. For this approach, simple derivations show that the ZTD estimator coincides with the tMLE or the G-computation estimator studied in Moore and van der Laan (2009a), as long as intercept terms are included. Moore and van der Laan (2009a) provide an excellent discussion of relationships among the estimating equation based, tMLE, and G-computation estimators for the dichotomous outcome problem. In addition, they provide an analytical relationship between the relative efficiency of these covariate-adjusted estimators and the predictive power of the covariates, which will help explain our simulation results presented later.

The augmented method in (7) is different from the usual regression method, i.e., fitting a logistic regression model including treatment assignment and baseline covariates, in two ways. First, in a logistic regression model, the parameter corresponding to the treatment effect is the log odds ratio conditional on the covariates in the model and in general is different from the overall (unconditional) log odds ratio, which has a one-to-one relationship to the VE parameter of interest. Therefore, whereas the usual regression method does not estimate VE, the augmented method does. Second, in a logistic regression model including treatment and baseline covariates, the estimation of the treatment effect is carried out simultaneously with the selection of covariates, raising concern over subjectivity due to the possibility of intentionally choosing the subset of covariates that best "accentuates" the treatment effect. The augmented method alleviates this concern by separating the regression of the response on covariates and the estimation of the treatment effect into two steps. In addition, in the regression step, because a separate model is built for each treatment group using only data for that group, two statisticians, who are blinded to the data in the other treatment group, could be assigned to build models for the two treatment groups, ensuring objectivity of the final estimation of treatment effect.

Instead of using separate models for $E(Y|Z = 1, X)$ and $E(Y|Z = 0, X)$, the augmented method may alternatively be implemented by specifying and fitting a model for $E(Y|Z, X)$, based on the data pooled over the the two treatment groups. This approach could be made objective by pre-specifying the model selection procedure in the study protocol. In addition, in some situations this approach will be more efficient in finite-samples than the approach that uses separate models. To see this, note that using separate models for the two treatment groups is equivalent to including all interactions of Z and X in the model for $E(Y|Z, X)$. If some of the interaction terms are not necessary, then this approach yields inefficient estimation of the nuisance parameters (i.e, the regression coefficients in the augmentation terms), and may also yield less efficient estimation of the treatment effect in finite samples (although not asymptotically). However, in addition to its strength to ensure an objective analysis, advantages of the separate-modeling approach include that it affords the data analyst the freedom to use any model building techniques, and it facilitates inclusion of higher order interaction or nonlinear terms that are supported by the data (and hence can improve efficiency), which may be hard to specify at the protocol development stage.

2.3. Estimating the variance of the augmented estimators. The augmented estimators (7) are M-estimators and therefore their asymptotic covariance matrix has the usual sandwich form, which may be estimated by the sandwich covariance estimator (Huber, 1967; Stefanski and Boos, 2002); see ZTD for details.

Specifically, the asymptotic covariance matrix for $\widehat{\beta}_{aug}$ is $\Delta^{-1}\Gamma(\Delta^{-1})^T$, where $\Delta = E\{-\frac{\partial m(Y,Z;\beta)}{\partial \beta^T}\}|_{\beta=\beta_T}$, $\Gamma = E[\{m^*_{aug}(Y,X,Z;\beta_T)\}^{\otimes 2}]$, $u^{\otimes 2} = uu^T$, and β_T is the true value of β. In practice, one may estimate the variance of $\widehat{\beta}_{aug}$ consistently by substituting all unknown quantities in the covariance matrix by the corresponding empirical parts, i.e.,

$$\widehat{\mathrm{var}}(\widehat{\beta}_{aug}) = n^{-1}\widehat{\Delta}^{-1}\Big[n^{-1}\sum_{i=1}^{n}\{m^*_{aug}(Y_i,Z_i,X_i;\widehat{\beta}_{aug})\}^{\otimes 2}\Big](\widehat{\Delta}^{-1})^T \tag{8}$$

where

$$\widehat{\Delta} = \begin{pmatrix} (1-\widehat{\pi})\frac{\exp(\widehat{\beta}_{aug0})}{\{1+\exp(\widehat{\beta}_{aug0})\}^2} & 0 \\ \widehat{\pi}\frac{\exp(\widehat{\beta}_{aug0}+\widehat{\beta}_{aug1})}{\{1+\exp(\widehat{\beta}_{aug0}+\widehat{\beta}_{aug1})\}^2} & \widehat{\pi}\frac{\exp(\widehat{\beta}_{aug0}+\widehat{\beta}_{aug1})}{\{1+\exp(\widehat{\beta}_{aug0}+\widehat{\beta}_{aug1})\}^2} \end{pmatrix}. \tag{9}$$

Tsiatis et al. (2007) studied the augmentation method in the setting of estimating the difference in means of two treatment groups and found that the sandwich covariance estimator may under-estimate the true variance when the sample size is small. They proposed a small-sample "correction factor," κ, such that the variance is estimated by κ times the usual sandwich covariance estimator, where

$$\kappa = \{(n_0 - p_0 - 1)^{-1} + (n_1 - p_1 - 1)^{-1}\}/\{(n_0 - 1)^{-1} + (n_1 - 1)^{-1}\}, \tag{10}$$

and p_0 and p_1 are the number of parameters in the models $q_0(X;\eta)$ and $q_1(X;\zeta)$, respectively, not including the intercept. We adopt this small-sample-size correction factor in our setting and further study its performance in simulation studies.

In the above presentation, we focused on estimating the log odds ratio β_1. To estimate VE, one may use $\widehat{VE} = 1 - \exp(\widehat{\beta}_1)$, whose variance can be estimated by the delta method; i.e,. $\widehat{\mathrm{var}}(\widehat{VE}) = \exp(2\widehat{\beta}_1)\widehat{\mathrm{var}}(\widehat{\beta}_1)$. Hypotheses about VE, such as $H_0 : VE \leq a$, can be tested using the equivalent form in terms of β_1, $H_0 : \beta_1 \geq \log(1-a)$. Confidence intervals for VE can be obtained by transforming the symmetric Wald-based confidence limits for β_1.

3 Extension to a Discrete Failure Time Outcome

In practice, prevention trials are often analyzed using discrete failure time methods. For example, in HIV prevention trials, HIV diagnostic tests are administered at fixed visits, and the discrete failure time is the visit-interval during which a subject is diagnosed. In this section, we discuss an extension of the semiparametric augmentation approach to data with a discrete failure time outcome. Let T_c denote the continuous failure time that falls into one of the intervals $[t_0 = 0, t_1), [t_1, t_2), \ldots, [t_{J-1}, t_J), [t_J, t_{J+1} = \infty)$, where t_J is the end of the study, and we may define the discrete failure as $T = t_j$ if $T_c \in [t_{j-1}, t_j)$. The distribution of the discrete failure time may be characterized by the discrete hazards at times $t_j, j = 1, \ldots, J$; the jth discrete hazard is defined as the conditional probability of failure at t_j given survival to that point, i.e.,

$$\lambda(t_j) = Pr(T = t_j | T > t_{j-1}). \tag{11}$$

The effect of treatment Z on failure time T may be modeled through the discrete hazards; a popular such model uses a logit link (Cox, 1972; Thompson, 1977),

$$\text{logit}\{\lambda(t_j|Z)\} = \alpha_j + \beta Z, \qquad j = 1, \ldots, J. \tag{12}$$

This model may be viewed as a logistic regression for each distinct time point t_j, $j = 1, \ldots, J$, where the odds ratio of failure for $Z = 1$ versus $Z = 0$ is assumed to be constant across the intervals $t_j, j = 1, \ldots, J$. When J is 1, for example, one is only interested in infection or not before the end of the study, and model (12) reduces to the usual logistic regression model.

As in almost all time-to-event trials, T is not fully observed due to early drop-out or to not failing by the end of study. With C the censoring time, we observe $V = \min(T, C)$ and $D = I(T \le C)$. Under the assumption of uninformative censoring, i.e., $T \perp\!\!\!\perp C | Z$, the parameters in (12) can be consistently estimated by solving the estimating equation

$$\sum_{i=1}^{n} h_i = \sum_{i=1}^{n} (h_{i1}, h_{i2}, \ldots, h_{iJ}, h_{i(J+1)})^T = 0, \tag{13}$$

where $h_{ij} = I(V_i \ge t_j)\{I(V_i = t_j, D_i = 1) - \text{expit}(\alpha_j + \beta Z_i)\}, j = 1, \ldots, J$, $h_{i(J+1)} = Z_i \sum_{j=1}^{J} I(V_i \ge t_j)\{I(V_i = t_j, D_j = 1) - \text{expit}(\alpha_j + \beta)\}$, and h_i is of dimension $(J+1)$. This method can be implemented using standard software for the logistic regression model by creating pseudo-observations d_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ with $m_i = \min_{j=1,\ldots,J}(V_i \le t_j)$, where $d_{ij} = I(V_i = t_j, D_i = 1)$.

Because the response T is censored for some subjects, the theory developed in ZTD does not directly apply to obtain more efficient estimators for β in (12). However, we notice that the key condition of independence between baseline covariates X and Z still holds, and motivated by the development by ZTD, we immediately deduce a class of unbiased estimating functions of the following form,

$$\sum_{i=1}^{n} h_i - (Z_i - \pi)a(X_i) = 0, \tag{14}$$

where $a(X_i)$ is a $(J+1)$ dimensional, arbitrary function of X_i. The optimal estimator in this class is the one that solves an estimating equation in this class with $a(X_i) = E(h_i|Z=1, X_i) - E(h_i|Z=0, X_i)$. However, unlike the class of estimators for the logistic regression model with dichotomous outcome characterized by (5), which includes all RAL estimators based on (Y, Z, X), the estimators characterized by (14) constitute only one class of estimators involving baseline covariates X, and do not include all RAL estimators for β based on (V, D, Z, X). Consequently, the optimal estimator is only optimal in this class and is not optimal among all possible reasonable estimators.

The solution to the estimating equation (14) does not have a closed form expression, and therefore we adopt the following approach:

(1) Use standard logistic regression software to obtain the usual unadjusted estimates of the h_i, for $i = 1, \ldots, n$.

(2) For each treatment group separately, fit linear regression models using OLS on baseline covariates for each of the first J components of h_i, and predict $E(h_i|Z=k, X_i)$ for all subjects, for each $k = 0, 1$.

(3) Substitute predicted values of $E(h_i|Z=1, X_i) - E(h_i|Z=0, X_i)$ for $a(X_i)$ into (14) and solve the equation to obtain the augmented estimator.

Note, in step two, we recommend fitting linear regression models using OLS and the rationale for doing so is explained in the next paragraph. The sandwich covariance estimator may be used to estimate the variance of the augmented estimator, and a small-sample bias correction factor κ may be applied, with $\kappa = \{(n_0 - p_0 - 1)^{-1} + (n_1 - p_1 - 1)^{-1}\}/\{(n_0 - 1)^{-1} + (n_1 - 1)^{-1}\}$, where p_0 is the average number of regressors (not including the intercept) of the J regression models for the placebo group, and p_1 is defined similarly for the treatment group.

It may seem non-standard to fit linear regression models for a dichotomous response variable m_i in the model for $E(m|Z=k, X)$ (in this case, equivalent to directly fitting linear models for $E(Y|Z=k, X)$ for dichotomous reponse Y) and for the trichotomous response variables h_{ij} in the discrete failure time models for $E(h_j|Z=k, X)$, $k = 0, 1, j = 1, \ldots J$. In theory, the optimal $a(X)$

among all functions of X is the corresponding conditional expectation, and therefore the proposed estimators are optimal in the corresponding class if the assumed model for $E(m|Z = k, X)$ or $E(h_j|Z = k, X)$, $j = 1, \ldots J$, contains the truth, which cannot be guaranteed in practice. As we may not necessarily expect to correctly specify the true functional forms of the regression models, especially for the multivariate response h in the discrete failure time model, a practical strategy is to restrict the search for optimal $a(X)$ within the class of functions linearly spanned by the basis functions specified in the regression models; see Leon et al. (2003) for a detailed explanation. According to Leon et al. (2003), if we fit the specified linear regression models by OLS, then the asymptotic variance of the corresponding augmented estimating function, e.g., $h - (Z - \pi)a(X)$ in (14), is guaranteed to be smaller than that of the unadjusted one, which corresponds to $a(X) = 0$ in the restricted class, and hence the corresponding augmented estimator is guaranteed to be more efficient. Therefore, the rationale for recommending fitting linear regression models for $E(h_j|Z = k, X)$ and $E(m|Z = k, X)$ using OLS is to minimize the variances of the augmented estimating equations; it is not our goal to make predictions for h or m.

The influence functions corresponding to class (14) belong to the class of all influence functions studied in Robins and Rotnitzky (1992) and Chapter 3 of van der Laan and Robins (2003) for right-censored data, and their general theory and results provided the theoretical basis for this development. In fact, according to van der Laan and Robins (2003), the estimating functions (14) can further be augmented by a term corresponding to the projection of h onto the nuisance tangent space of censoring, and this leads to the optimal estimating function for the fixed h. In (14), the fixed h is one of many possible choices of unadjusted estimating functions (or influence functions). In the two references given above and Robins (1993), the form of the optimal unadjusted estimating function is derived and methods for numerically obtaining an estimate of it are studied. This together with the second augmentation term will lead to the optimal or locally semiparametric efficient estimator among all possible RAL estimators. We refer the readers to van der Laan and Robins (2003) for technical details and a comprehensive list of references where augmented inverse probability weighting or locally efficient estimators for general right-censored data are studied. In particular, for randomized trials with right-censored outcome, Hubbard, van der Laan, and Robins (1999) studied the locally efficient estimator for the marginal survival probability at a fixed time-point based on solving estimating equations. In this case, as there is only one unadjusted influence function, one does not need to numerically estimate the optimal one and augmenting any unadjusted estimating function leads to the locally efficient estimator. Alternatively, Moore and van der Laan (2009b) studied the tMLE for the same parameter, where an iterative procedure is used to update an initial fit of the likelihood, and this estimator also solves the efficient estimating equation.

In our proposed method, to ease computation we fix h as the one commonly used in unadjusted analysis, and we do not include a second augmentation term, which would require additional modeling effort. While significantly reducing the complexity of implementation, this simplification comes at the price of not achieving the best possible efficiency gain. It would be of interest to study the locally efficient estimator for β in (12) in the setting of randomized trials and to evaluate the trade-off between computational complexity and efficiency, which we leave as future work.

4 Simulation Study

We conduct a simulation study to evaluate the performance of the methods for addressing the primary objective in a Phase 2b prevention trial. For concreteness we simulate a prototype Phase 2b HIV vaccine efficacy trial with objective to evaluate VE. All reported results are based on 5000 simulated data sets.

4.1. Simulation study for dichotomous outcome. ZTD developed the asymptotic properties of the augmented estimator considered in Section 3 and illustrated in simulations that considerable efficiency gains may be achieved when baseline covariates are moderately correlated with the dichotomous outcome. In this section, we further study this estimator in simulations of a rare event prevention trial, to evaluate (1) the degree of efficiency gain when covariates are only mildly correlated with the outcome; and (2) the validity and efficiency when the sample size is moderate (60 to 120 total events). We study scenario (1) because it often occurs in Phase 2b and 3 prevention trials, and (2) because it occurs in Phase 2b prevention trials.

Our prototype trial has an event rate of approximately 10% in the placebo group, reflecting the typical rare event rate of HIV infection in an HIV prevention trial. We choose the true VE to be around 0.35 (equivalently, a log odds ratio of -0.4). We generate the treatment assignment Z as Bernoulli with $\pi = 0.5$, and, independent of Z, generate a vector of baseline covariates $X = (X_1, \ldots, X_{20})^T$, where X_1 to X_3 and X_5 to X_8 are normal random variables and the rest are binary. Correlations among the variables are created such that some "important"covariates (defined shortly) are correlated with "unimportant" covariates, and normally distributed covariates are correlated with binary covariates. The binary outcome Y is generated as Bernoulli conditional on X within each treatment group according to $\text{logit}\{P(Y = 1|Z = k, X)\} = a_{0k} + a_{1k}^T X_{imp}$, $k = 0$ or 1, where $X_{imp} = (X_1, \ldots, X_4,)^T$, and are referred to as "important" covariates. We choose coefficients a_{0k} and a_{1k} such that the coefficient of determination, R^2, as a measure of the strength of associations between covariates and outcome within each treatment group, is about 0.05, 0.10 or 0.15. Sample sizes of $n = 750$ and $n = 1500$ are considered for each of these three scenarios.

We estimate the log odds ratio β_1 using the standard MLE and using seven augmented estimators, which develop regression models for each treatment group via the following approaches:

Aug. 1: fit a logistic regression model using the true covariates (the true model).
Aug. 2: fit a logistic regression model using all X.
Aug. 3: fit a linear regression model using true covariates by OLS.
Aug. 4: fit a linear regression model using all X by OLS.
Aug. 5: build a logistic regression model with all X and select covariates by forward selection with entry level equal to 0.05.
Aug. 6: build a logistic regression model with all X and the square of all continuous covariates and select covariates by forward selection with entry level equal to 0.05.
Aug. 7: build a logistic regression model with covariates $(X_1, X_4, X_7)^T$ and $(X_1, X_4, X_5, X_{20})^T$ for the placebo and vaccine group respectively (a wrong model).

In addition to examining bias of the eight estimators of β_1, we investigate performance of the sandwich variance estimator with and without the small-sample correction. This informs about the effectiveness of the small-sample correction and about the relative efficiency of the estimators. In addition, we examine coverage probability of the Wald-based confidence intervals for β_1, which equivalently examines coverage probability of the transformed symmetric confidence intervals for VE. Finally, we implement the bootstrap variance estimator for scenario 3 when the sample size is 750. Due to computational constraints, the bootstrap variance estimators in the simulations are based on 50 bootstrap samples.

The results are reported in Tables 2-4. All of the estimators have small biases and accurate coverage probabilities. The augmented estimators are more efficient, about 2%, 7% and 15% so for scenarios 1, 2 and 3. The sandwich variance estimators exhibit only slight underestimation without the correction factor, which is mostly corrected by the correction factor. The bootstrap standard errors are slightly more accurate. The results for the model that misspecified the augmentation term (Aug. 7) show that the misspecification we studied did not lead to poor performance. However, it is possible for a badly misspecified model to lead to reduced power in finite samples.

Table 2: *Estimation of the log-odds ratio: Scenario 1. "Standard" refers to the maximum likelihood estimator and the augmented estimators are described in Section 4. MC bias is the Monte Carlo bias, MC SD is Monte Carlo Standard deviation, SE is the average of estimated standard errors, CP is the Monte Carlo coverage probability of* 95% *Wald confidence intervals, C. SE is the average of estimated standard errors corrected by the small-sample-size correction factor, C. CP is the Monte Carlo coverage probability of* 95% *Wald confidence intervals using the corrected SE, and RE is relative efficiency calculated as the Monte Carlo mean squared error for the standard estimator divided by that for the indicated estimator.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	RE
		n=1500, truth=-0.457					
Standard	-0.008	0.190	0.189	0.953			1
Aug. 1	-0.008	0.187	0.185	0.955	0.185	0.955	1.033
Aug. 2	-0.007	0.188	0.183	0.951	0.185	0.953	1.018
Aug. 3	-0.008	0.187	0.186	0.954	0.186	0.954	1.029
Aug. 4	-0.008	0.188	0.185	0.950	0.187	0.954	1.018
Aug. 5	-0.008	0.187	0.184	0.952	0.185	0.953	1.027
Aug. 6	-0.008	0.188	0.184	0.950	0.185	0.950	1.024
Aug. 7	-0.008	0.187	0.186	0.952	0.186	0.953	1.027
		n=750, truth=-0.457					
Standard	-0.012	0.270	0.269	0.954			1
Aug. 1	-0.011	0.267	0.263	0.953	0.264	0.954	1.028
Aug. 2	-0.005	0.269	0.256	0.944	0.263	0.950	1.014
Aug. 3	-0.012	0.267	0.264	0.953	0.265	0.954	1.029
Aug. 4	-0.013	0.270	0.262	0.948	0.269	0.957	1.004
Aug. 5	-0.011	0.269	0.262	0.949	0.263	0.950	1.014
Aug. 6	-0.011	0.269	0.261	0.950	0.262	0.952	1.014
Aug. 7	-0.012	0.267	0.264	0.956	0.266	0.956	1.024

Supplementary Tables 1-3 show parallel results for estimating VE, showing that there is more finite-sample bias for estimating VE than β_1; this occurs because $\widehat{VE}$ has a less symmetrical distribution than $\widehat{\beta}_1$.

4.2. Simulation study for discrete failure time outcome.

For a discrete failure time outcome, we consider a 1000-subject trial with six evenly spaced follow-up visits, at which HIV diagnostic tests are administered. We generate the treatment assignment Z as Bernoulli with $\pi = 0.5$, and generate four baseline covariates $X = (X_1, \ldots, X_4)^T$, independent of treatment

Table 3: *Estimation of log-odds ratio: Scenario 2. Entries are as in Table 2.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	RE
			n=1500, truth=-0.410				
Standard	-0.007	0.189	0.188	0.951			1
Aug. 1	-0.006	0.180	0.179	0.952	0.179	0.952	1.099
Aug. 2	-0.005	0.181	0.177	0.948	0.179	0.952	1.086
Aug. 3	-0.007	0.183	0.182	0.951	0.182	0.952	1.070
Aug. 4	-0.007	0.183	0.181	0.950	0.183	0.954	1.061
Aug. 5	-0.006	0.181	0.178	0.949	0.179	0.950	1.094
Aug. 6	-0.006	0.181	0.178	0.949	0.179	0.950	1.090
Aug. 7	-0.007	0.183	0.182	0.953	0.182	0.953	1.068
			n=750, truth=-0.410				
Standard	-0.012	0.271	0.268	0.953			1
Aug. 1	-0.010	0.259	0.254	0.950	0.255	0.950	1.090
Aug. 2	-0.004	0.263	0.247	0.938	0.254	0.945	1.060
Aug. 3	-0.012	0.262	0.259	0.951	0.260	0.952	1.065
Aug. 4	-0.012	0.265	0.256	0.944	0.263	0.950	1.039
Aug. 5	-0.010	0.262	0.253	0.944	0.254	0.945	1.070
Aug. 6	-0.009	0.262	0.252	0.944	0.253	0.945	1.068
Aug. 7	-0.012	0.263	0.259	0.950	0.260	0.951	1.060

Table 4: *Eestimation of the log-odds ratio: Scenario 3. B. SE is Bootstrap standard error, B. CP is Bootstrap confidence interval and other entries are as in Table 2.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	B. SE	B. CP	RE
				n=1500, truth=-0.397					
Standard	-0.005	0.187	0.185	0.951					1
Aug. 1	-0.005	0.171	0.168	0.950	0.169	0.950			1.196
Aug. 2	-0.003	0.172	0.166	0.944	0.168	0.946			1.182
Aug. 3	-0.005	0.176	0.174	0.951	0.175	0.951			1.125
Aug. 4	-0.005	0.177	0.173	0.946	0.176	0.949			1.111
Aug. 5	-0.005	0.171	0.168	0.950	0.168	0.950			1.192
Aug. 6	-0.004	0.171	0.168	0.950	0.168	0.951			1.189
Aug. 7	-0.006	0.176	0.174	0.950	0.174	0.950			1.125
				n=750, truth=-0.397					
Standard	-0.008	0.270	0.263	0.946			0.267	0.943	1
Aug. 1	-0.005	0.248	0.239	0.942	0.240	0.943	0.245	0.943	1.187
Aug. 2	0.002	0.253	0.231	0.930	0.238	0.939	0.246	0.941	1.135
Aug. 3	-0.008	0.255	0.248	0.945	0.249	0.946	0.252	0.944	1.122
Aug. 4	-0.008	0.258	0.245	0.941	0.252	0.948	0.256	0.946	1.092
Aug. 5	-0.005	0.252	0.238	0.939	0.239	0.939	0.251	0.947	1.153
Aug. 6	-0.003	0.253	0.237	0.936	0.238	0.937	0.254	0.951	1.144
Aug. 7	-0.008	0.256	0.247	0.943	0.248	0.945	0.252	0.945	1.115

assignment Z, where X_4 is Bernoulli (0.5), X_1, X_2, X_3 are marginally standard normal, and covariates are independent of each other except that X_1 and X_2 are correlated with correlation coefficient 0.2. The discrete failure time follows a geometric distribution with parameters 0.02 and 0.0133 for $Z = 0$ and $Z = 1$, so that $VE \approx 0.35$. Specifically, we first generate $U = \Phi\{(\eta X_1 + \epsilon)/\sqrt{\eta^2 + 1}\}$, where $\epsilon \sim N(0, 1)$ and Φ is the cumulative distribution function (cdf) of the standard normal distribution, and hence U is marginally uniform (0,1). Then we generate $T|Z$ by transforming U using the inverse cdf of the geometric distribution. This two-step procedure creates an association between T and X_1 and in the meantime guarantees that the distribution of T given Z follows the intended distribution. The censoring time C is generated from a mixture distribution of binomial(5, 0.5) with probability 0.2 and constant 6 with probability 0.85. Therefore, most of the subjects are censored after the sixth follow-up interval; the event rate is about 10% in the placebo group.

Data were generated under three scenarios, wherein η was chosen so that the correlation coefficient between the possibly censored event time and X_1 is about 0.10, 0.15 or 0.20 for each treatment group. In addition to the usual estimator that ignores baseline covariates, we implement four versions of the augmented estimator, which develop regression models for each treatment group via the following approaches:

Aug. 1: fit a linear regression model using X_1 by OLS.

Aug. 2: fit a linear regression model using X_1 and the square of X_1 by OLS.

Aug. 3: fit a linear regression model using all X, square and interaction terms by OLS.

Aug. 4: build a linear regression model using all X, square and interaction terms and select covariates by forward selection with entry level equal to 0.15.

The results are reported in Table 5. Both the standard estimator and the augmented estimators perform well (unbiased and correct coverage probabilities), with the augmented estimators more efficient than the standard estimator (about 5%, 10% and 20% more efficient for scenarios 1, 2 and 3).

5 Application to the First HIV Vaccine Efficacy Trial

We assess VE in the North America/Netherlands VaxGen HIV vaccine efficacy trial using the methods evaluated above, first treating HIV infection as a

Table 5: *Estimation of β in (12) for discrete failure time outcome. Entries are as in Table 2.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	RE
			n=1000, truth=-0.415				
			Discrete failure time, scenario 1				
Standard	-0.011	0.224	0.222	0.949			1
Aug. 1	-0.011	0.219	0.217	0.950	0.217	0.950	1.048
Aug. 2	-0.010	0.218	0.216	0.949	0.216	0.950	1.058
Aug. 3	-0.010	0.220	0.215	0.946	0.218	0.949	1.040
Aug. 4	-0.010	0.219	0.215	0.944	0.216	0.945	1.043
			Discrete failure time, scenario 2				
Standard	-0.010	0.224	0.222	0.953			1
Aug. 1	-0.010	0.214	0.211	0.948	0.212	0.949	1.094
Aug. 2	-0.009	0.210	0.207	0.950	0.208	0.950	1.137
Aug. 3	-0.009	0.212	0.206	0.946	0.209	0.950	1.113
Aug. 4	-0.009	0.212	0.207	0.949	0.208	0.950	1.121
			Discrete failure time, scenario 3				
Standard	-0.009	0.223	0.222	0.958			1
Aug. 1	-0.009	0.208	0.205	0.955	0.206	0.955	1.151
Aug. 2	-0.007	0.199	0.196	0.953	0.197	0.953	1.251
Aug. 3	-0.007	0.202	0.195	0.948	0.198	0.952	1.222
Aug. 4	-0.007	0.201	0.196	0.950	0.197	0.951	1.227

dichotomous endpoint and secondly as a discrete failure time endpoint. Applying the logistic regression model (1), the MLE of the log odds ratio of infection for vaccine versus placebo is -0.053 (SE: 0.114; 95% CI: -0.276 – 0.170) and the MLE of VE is 0.052 (SE: 0.108; 95% CI: -0.189 – 0.241). We next applied the augmented method, where treatment-specific regression models are developed to predict HIV infecion as a dichotomous endpoint using the forward selection method with entry level p-value of 0.25 and baseline variables **riskscore** (an integer index taking values 0–7 measuring the number of independent risk factors for HIV infection, which was derived from multivariate Cox modeling; Flynn et al., 2005); **sex**; race (white, black, hispanic, or asian), country of residence (**U.S.**, Canada, or Netherlands), sexual risk behaviors in the past 6 months (number of sex partners, number of HIV+ sex partners, number of male sex partners, number of male HIV+ sex partners, number of male HIV- partners, number of male HIV status unknown sex partners); sexually transmitted infections (hepititis B, chlamydia, **gonorrhea**, warts); and drug use in the past 6 months (alcohol, amphetamines, crack, hallucinogens, marijuana, poppers, **tranquilizers**). For the placebo group, the selected covariates are white, black, hispanic, number of male HIV status unknown sex partners, heptititis B, chlamydia, alcohol, amphetamines, crack and the five bolded variables. For the vaccine group, the selected variables include the bolded variables, number of male HIV+ sex partners, number of male HIV- partners, warts, hallucinogens, and poppers.

The augmented estimate of the log odds ratio is -0.050 (SE: 0.111; 95% CI: -0.267 – 0.168) and of the VE is 0.049 (SE: 0.106; 95% CI: -0.182 – 0.235). The efficiency of the augmented estimator relative to the standard estimator is 1.049. Using the small sample correction factor, the estimated standard error for the log odds ratio estimate is 0.111 (95% CI: -0.268 – 0.168), and for estimating VE, the estimated standard error is 0.106 (95% CI: -0.183 – 0.235). The efficiency relative to the standard estimator is 1.043, calculated using the corrected standard errors. Additional results are reported in Table 6. Bootstrap standard error estimates are based on 1000 bootstrap samples.

We next analyzed the data using a discrete failure time outcome that takes values in the monthly intervals $[0, 6), [6, 12), [12, 18), [18, 24), [24, 30), [30, 36)$. Applying the standard method (12), the estimated common log odds ratio is -0.061 (SE: 0.110; 95% CI: -0.278 – 0.155; Bootstrap SE: 0.112), and the corresponding VE estimate is 0.060 (SE: 0.104; 95% CI: -0.168 – 0.243; Bootstrap SE: 0.107). Applying the proposed augmented method with regression models built using forward model selection models with entry level p-value of 0.25, the estimated common log odds ratio is -0.051 (SE: 0.108; 95% CI: -0.263 – 0.161; Bootstrap SE: 0.110) and of the VE is 0.050 (SE: 0.103; 95% CI: -0.175 – 0.232; Bootstrap: 0.106). The efficiency of the augmented estimator relative to the standard estimator, based on analytical standard error estimates, is 1.043. Using the small sample correction factor, the standard error for the augmented estimator of the log odds ratio is 0.108 (95% CI: is -0.264 – 0.161) and the efficiency relative to the standard estimator is 1.041.

To mimic a Phase 2b trial, we next took a random sample of the VaxGen data, with each subject selected with probability 100/368. Results based on a binary outcome are shown in Table 7. For a discrete failure time outcome, the estimated common log odds ratio using the standard method is 0.184 (SE: 0.214; 95% CI: -0.236 – 0.604; Bootstrap SE: 0.230). Applying the augmented method with regression models built using forward model selection method with entry level p-value of 0.25, the estimated common log odds ratio is 0.152 (SE: 0.203; 95% CI: -0.246 – 0.549; Bootstrap SE: 0.231) and the efficiency relative to the standard estimator, based on analytical standard error estimates, is 1.116. Using the small sample correction factor, the standard error for the augmented estimator of the log odds ratio is 0.203 (95% CI: -0.247 – 0.550) and the efficiency relative to the standard estimator is 1.110.

Table 6: *Data analysis of North America/Netherlands VaxGen HIV vaccine efficacy trial: Forward selection with entry level equal to sle was used to select regression models for augmented method;* R^2 *is the coefficient of determination in the corresponding regression model; Est is the point estimate; SE is the estimated standard error; C. SE is the estimated corrected standard error; B. SE is the Bootstrap standard error; RE, C. RE, and B. RE are estimated relative efficiency, calculated as the square of ratios of the three standard error estimates for the corresponding estimator with the standard estimator as reference.*

					log odds ratio					VE		
Method	R^2	Est	SE	C. SE	B. SE	RE	C. RE	B. RE	Est	SE	C. SE	B. SE
Standard		-0.053	0.114		0.115	1	1	1	0.0515	0.108		0.111
Aug. 1 (sle=0.05)	0.057 (Z=0) 0.028 (Z=1)	-0.440	0.112	0.112	0.113	1.036	1.034	1.023	0.043	0.107	0.107	0.110
Aug. 2 (sle=0.15)	0.060 (Z=0) 0.032 (Z=1)	-0.478	0.111	0.111	0.113	1.043	1.039	1.025	0.047	0.106	0.106	0.110
Aug. 3 (sle=0.25)	0.068 (Z=0) 0.033 (Z=1)	-0.050	0.111	0.111	0.113-0	1.049	1.043	1.023	0.049	0.106	0.106	0.110

Table 7: *Data analysis of VaxGen HIV vaccine efficacy trial: A random subsample of sample size 1429. Entries are as in Table 6.*

					log odds ratio					VE		
Method	R^2	Est	SE	C. SE	B. SE	RE	C. RE	B. RE	Est	SE	C. SE	B. SE
Standard		0.197	0.221		0.227	1	1	1	-0.218	0.268		0.291
Aug. 1 (sle=0.05)	0.077 (Z=0) 0.050 (Z=1)	0.192	0.214	0.215	0.227	1.069	1.064	1.007	-0.212	0.259	0.260	0.281
Aug. 2 (sle=0.15)	0.093 (Z=0) 0.052 (Z=1)	0.192	0.213	0.214	0.227	1.078	1.067	1.006	-0.212	0.258	0.260	0.279
Aug. 3 (sle=0.25)	0.110 (Z=0) 0.056 (Z=1)	0.188	0.212	0.213	0.228	1.092	1.076	0.997	-0.206	0.256	0.257	0.280

DOI: 10.2202/1948-4690.1002

6 Discussion

We evaluated by simulation the finite-sample performance of the semiparametric efficient method of ZTD for estimating the log-odds-ratio or equivalently VE for a dichotomous endpoint, and proposed a new semiparametric method for incorporating baseline covariates for estimating the common log-odds-ratio or VE in a discrete failure time model. The relative efficiency of the semiparametric method compared to the standard unadjusted estimator (that does not account for baseline covariates) for both dichotomous and discrete failure time outcome models were evaluated in settings that mimic Phase 2b prevention trials and for which the baseline covariates are only mildly correlated with the outcome. We also assessed a small-sample correction factor used for correcting bias in the variance estimators. We applied the methods to the first HIV vaccine efficacy trial.

Although consistency and asymptotic normality of the semiparametric estimators are based on large-sample theory, our simulation study suggests their appropriate use in intermediate-sized Phase 2b efficacy trials as, in all simulated scenarios, the estimators are approximately unbiased, and the confidence intervals have close to nominal coverage levels, with the small-sample correction factor improving coverage. Moreover, the semiparametric method achieves approximately 5% to 15% gains in efficiency when covariates are only mildly correlated with outcome. Application of the method to the North America/Netherlands VaxGen HIV vaccine efficacy trial demonstrates that such improvement in efficiency is possible with real data.

Appendix

Supplementary Tables 1: *Estimation of VE: Scenario 1. Entries are as in Table 2.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	RE
			n=1500, truth=0.367				
Standard	-0.006	0.122	0.123	0.952			1
Aug. 1	-0.006	0.120	0.118	0.946	0.118	0.946	1.035
Aug. 2	-0.007	0.121	0.117	0.941	0.118	0.945	1.016
Aug. 3	-0.006	0.120	0.118	0.946	0.118	0.946	1.030
Aug. 4	-0.006	0.121	0.118	0.942	0.119	0.944	1.019
Aug. 5	-0.006	0.120	0.118	0.943	0.118	0.944	1.028
Aug. 6	-0.006	0.120	0.117	0.943	0.118	0.943	1.024
Aug. 7	-0.006	0.120	0.118	0.945	0.119	0.945	1.029
			n=750, truth=0.367				
Standard	-0.016	0.178	0.177	0.946			1
Aug. 1	-0.016	0.175	0.169	0.937	0.170	0.938	1.035
Aug. 2	-0.020	0.177	0.166	0.931	0.171	0.936	1.003
Aug. 3	-0.015	0.175	0.170	0.937	0.171	0.938	1.036
Aug. 4	-0.015	0.177	0.168	0.930	0.173	0.937	1.014
Aug. 5	-0.016	0.176	0.169	0.934	0.169	0.935	1.024
Aug. 6	-0.016	0.176	0.168	0.933	0.169	0.934	1.024
Aug. 7	-0.015	0.175	0.170	0.937	0.171	0.937	1.030

Supplementary Tables 2: *Estimation of VE: Scenario 2. Entries are as in Table 2.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	RE
			n=1500, truth=0.337				
Standard	-0.008	0.127	0.128	0.950			1
Aug. 1	-0.007	0.121	0.120	0.944	0.120	0.944	1.103
Aug. 2	-0.008	0.122	0.118	0.941	0.120	0.944	1.084
Aug. 3	-0.007	0.123	0.121	0.944	0.122	0.945	1.073
Aug. 4	-0.007	0.123	0.121	0.942	0.122	0.943	1.064
Aug. 5	-0.007	0.121	0.119	0.944	0.120	0.944	1.099
Aug. 6	-0.007	0.122	0.119	0.942	0.119	0.942	1.094
Aug. 7	-0.007	0.123	0.121	0.943	0.122	0.944	1.072
			n=750, truth=0.337				
Standard	-0.016	0.185	0.185	0.943			1
Aug. 1	-0.015	0.177	0.172	0.935	0.172	0.937	1.097
Aug. 2	-0.021	0.181	0.168	0.928	0.173	0.933	1.044
Aug. 3	-0.015	0.179	0.174	0.935	0.175	0.936	1.076
Aug. 4	-0.015	0.181	0.173	0.929	0.178	0.935	1.050
Aug. 5	-0.016	0.179	0.171	0.933	0.172	0.934	1.075
Aug. 6	-0.017	0.179	0.171	0.932	0.171	0.933	1.070
Aug. 7	-0.015	0.179	0.174	0.933	0.175	0.934	1.069

Supplementary Tables 3: *Estimation of VE: Scenario 3. Entries are as in Table 4.*

Method	MC Bias	MC SD	SE	CP	C. SE	C. CP	B. SE.	C. SE	RE
				n=1500, truth=0.328					
Standard	-0.008	0.127	0.128	0.950					1
Aug. 1	-0.007	0.116	0.114	0.943	0.114	0.944			1.204
Aug. 2	-0.008	0.117	0.113	0.940	0.114	0.944			1.186
Aug. 3	-0.007	0.120	0.118	0.942	0.118	0.942			1.131
Aug. 4	-0.007	0.120	0.117	0.939	0.119	0.943			1.119
Aug. 5	-0.007	0.116	0.114	0.942	0.114	0.942			1.201
Aug. 6	-0.007	0.116	0.114	0.942	0.114	0.942			1.197
Aug. 7	-0.006	0.119	0.118	0.943	0.118	0.944			1.133
				n=750, truth=0.328					
Standard	-0.019	0.190	0.184	0.946			0.191	0.939	1
Aug. 1	-0.017	0.173	0.164	0.936	0.165	0.937	0.173	0.940	1.206
Aug. 2	-0.023	0.179	0.160	0.926	0.165	0.933	0.180	0.948	1.124
Aug. 3	-0.016	0.178	0.170	0.936	0.171	0.937	0.178	0.940	1.137
Aug. 4	-0.017	0.180	0.168	0.933	0.173	0.939	0.182	0.939	1.110
Aug. 5	-0.018	0.175	0.163	0.933	0.164	0.934	0.179	0.946	1.170
Aug. 6	-0.019	0.176	0.163	0.931	0.164	0.932	0.183	0.949	1.156
Aug. 7	-0.017	0.179	0.169	0.938	0.170	0.938	0.178	0.939	1.131

References

[1] Buchbinder, S. P., Mehrotra, D. V., Duerr, A., Fitzgerald, D. W., Mogg, R., Li, D., Gilbert, P. B., Lama, J. R., Marmor, M., del Rio, C., McElrath, M. J., Casimiro, D. R., Gottesdiener, K. M., Chodakewitz, J. A., Corey, L., Robertson, M. N. (2008). The Step study: The first test-of-concept efficacy trial of a cell-mediated immunity HIV vaccine. *Lancet* **372,** 1881-1893.

[2] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Boca Raton: Chapman and Hall/CRC.

[3] Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

[4] Cuzick, J. (1982). The efficiency of the proportions test and the logrank test for censored survival data. *Biometrics* **38,** 1033-1039.

[5] Fleming, T. R., Richardson, B. A. (2004). Some design issues in trials of microbicides for the prevention of HIV infection. *The Journal of Infectious Diseases* **190,** 666–674.

[6] Flynn, N. M., Forthal, D. N., Harro, C. D., Judson, F. N., Mayer, K. H., Para, M. F., and the rgp120 HIV Vaccine Study Group (2005). Placebo-controlled phase 3 trial of recombinant glycoprotein 120 vaccine to prevent HIV-1 infection *The Journal of Infectious Diseases* **191,** 654–665.

[7] Gilbert, P. B. (2009). Some design issues in Phase 2b versus Phase 3 prevention trials for testing efficacy of products or concepts. *Statistics in Medicine*, in press.

[8] Huber, P. J. (1967). The behavior of the maximum likelihood estimator under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 221–233.

[9] Leon, S. Tsiatis, A. A., and Davidian, M. (2003). Semiparametric efficient estimation of treatment effect in a pretest-posttest study. *Biometrics* **59,** 1046–1055.

[10] Lu X, Tsiatis, A. A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika* **95,** 679-694.

[11] Moore, K. L. and van der Laan, M. J. (2009a). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine* **28(1),** 39–64.

[12] Moore, K. L. and van der Laan, M. J. (2009b). Application of Time-to-Event Methods in the Assessment of Safety in Clinical Trials (Chapter 20; 455–482). *Design and Analysis of Clinical Trials with Time-to-Event Endpoints* [Peace, KE, ed.] Chapman and Hall/CRC Biostatistics Series.

[13] Pitisuttithum, P., Gilbert, P. B., Gurwith, M., Heyward, W., Martin, M., van Griensven, F., Hu, D., Tappero, J. W., Choopanya, K., and the Bangkok Vaccine Evaluation Group. (2006). Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. *The Journal of Infectious Diseases* **194,** 1661–1671.

[14] Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., Premsri, N., Namwat, C., de Souza, M., Adams, E., Benenson, M., Gurunathan, S., Tartaglia, J., McNeil, J.G., Francis, D.P., Stablein, D., Birx, D.L., Chunsuttiwat, S., Khamboonruang, C., Thongcharoen, P., Robb, M.L., Michael, N.L., Kunasol, P., Kim, J.H.; MOPH-TAVEG Investigators (2009). Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine* **361**:2209–2220.

[15] Rida, W., Fast, P., Hoff, R., Fleming, T. R. (1997). Intermediate-size trials for the evaluation of an HIV vaccine candidate: a workshop summary. *Journal of the Acquired Immune Deficiency Syndrome and Human Retrovirology* **16,** 195–203.

[16] Robins, J, M. (1993) Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the Biopharmaceutical section, American Statistical Association*, American Statistical Association, Alexandria, VA, 24-33.

[17] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

[18] Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statis- tician* **56,** 29–38.

[19] Thompson Jr., W. A. (1977) On the treatment of grouped observations in life studies. *Biometrics* **33,** 463–470.

[20] Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2007) Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27,** 4658–4677.

[21] van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, New York.

[22] van der Laan M. J. and Rubin D. (2006) Targeted maximum likelihood learning. *The International Journal of Biostatistics* Vol 2, Iss. 1, Article 11.

[23] Zhang, M., Tsiatis, A. A., and Davidian, M. (2008) Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64,** 707–715.