# Methods for Comparing Durability of Immune Responses Between Vaccine Regimens in Early-Phase Trials

Ted Westling[1], Michal Juraska[2], Kelly Seaton[3], Georgia Tomaras[3], Peter Gilbert[2], and Holly Janes[2]

[1]University of Washington, Seattle, WA, USA
[2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[3]Duke University, Durham, NC, USA

**Abstract**

The ability to produce a long-lasting, or durable, immune response is a crucial characteristic of many highly effective vaccines. A goal of early-phase vaccine trials is often to compare the immune response durability of multiple tested vaccine regimens. One parameter for measuring immune response durability is the area under the mean post-peak log immune response profile. In this paper, we compare immune response durability across vaccine regimens within and between two phase I trials of DNA-primed HIV vaccine regimens, HVTN 094 and HVTN 096. We compare four estimators of this durability parameter and the resulting statistical inferences for comparing vaccine regimens. Two of these estimators use the trapezoid rule as an empirical approximation of the area under the marginal log response curve, and the other two estimators are based on linear and nonlinear models for the marginal mean log response. We conduct a simulation study to compare the four estimators, provide guidance on estimator selection, and use the nonlinear marginal mean model to analyze immunogenicity data from the two HIV vaccine trials.

## 1 Introduction

Early-phase (phase I/IIa) clinical trials of candidate preventative vaccines are typically designed to evaluate immune responses that are generated by the tested vaccine, in addition to vaccine safety and tolerability. Immune responses usually peak shortly after the vaccination series is completed and wane over time. Evaluating the peak immune response is typically a primary objective. A secondary objective is often to evaluate immune response durability, or how long immune responses last. A key challenge for many pathogens is developing vaccines that generate durable responses. In particular, all previously tested HIV vaccines that were designed to elicit a humoral immune response generated antibody responses that declined in the majority of trial participants within one to three years following the last vaccination[1–5]. The phase III RV144 trial of an ALVAC/AIDSVAX prime/boost vaccine regimen[6] versus placebo supports the importance of immune response durability: vaccine efficacy against HIV-1 acquisition waned over time (60.5% at 12 months post-first vaccination but only 31.2% at 42 months post-first vaccination), as did the anti-envelope V1V2 IgG antibody response that correlated with decreased HIV-1 risk and was hypothesized to be partially responsible for protection[1,3,7] – suggesting that increasing the durability of this immune response could help

preserve vaccine efficacy. Indeed, for any vaccine to confer protection from infection or disease long after vaccinations are completed, it is important that the immune response elicited be not only protective, but also durable.

A typical early-phase vaccine trial randomizes healthy, uninfected participants to one of potentially multiple vaccine regimens, or placebo. Specimens are collected during the vaccination series, at the presumed time point of peak immune response (typically shortly after the last vaccination), and at a handful of fixed time points thereafter. Regimens are selected for further evaluation based on the immune responses elicited shortly after the final vaccination - at the presumed peak time point- and the durability of these immune responses. Ultimately phase IIb/III trials are conducted to evaluate preventative efficacy. While comparisons of regimens based on peak responses can rely on standard $k$-group comparisons of univariate outcomes, there are currently no standard statistical methods for evaluating or comparing the durability of vaccine-induced immune responses. One simple approach is to compare the distribution of immune response at a single late time point, e.g. the terminal visit, but a measure that makes use of all the immunogenicity data following the post-presumed peak time point may be preferable because such a measurement would better capture the kinetics of response waning. For example, suppose two vaccines induce immune responses that wane at different rates after the peak response, but both have undetectable immune responses by the terminal visit. If only presumed peak response and terminal visit data are analyzed, it is not possible to conclude which of these two vaccines elicits a more durable response. If, however, intermediate time points are included in the analysis as well, this problem may be circumvented. Additionally, an analysis that uses all the post-presumed peak data will be more statistically efficient. This is especially important in early-phase trials, which frequently have few participants per regimen.

Our work on this problem is motivated by data generated from two phase I HIV vaccine trials, HVTN 094 (ClinicalTrials.gov identifier NCT01571960; Buchbinder et al., 2017) and HVTN 096 (ClinicalTrials.gov identifier NCT01799954;[8]). HVTN 094 assigned fifteen participants per regimen and HVTN 096 assigned twenty per regimen. A key objective of each trial was to compare regimens with respect to durability of the immune responses generated by the vaccine. Because the vaccines tested in the two studies were considerably different in their make-up, it is also of interest to compare regimens across studies.

In this article, we illustrate the process of selecting a statistical method for contrasting the durability of immune responses between vaccine regimens using the HVTN 094 and HVTN 096 data. We measure immune response durability using the area under the marginal mean log immune response curve (AUC) after a presumed "peak response" time point. Three of the estimation methods we study are common, generic methods that can be used with almost any type of immunological assay data. The fourth method we use is based on a parametric model for the marginal mean immune response curve, selected to capture

the nonlinear decay of the immune responses in HVTN 094 and HVTN 096. We conduct a simulation study based on the HVTN 094 and HVTN 096 data to assess whether the selected parametric model outperforms the other methods, and which of the remaining three common estimators performs best.

In the next section we describe the HVTN 094 and HVTN 096 data. In Section 3 we define the four estimators that we consider. In Section 4 we describe the results of our simulation study. In Section 5 we present the results of our analyses of HVTN 094 and HVTN 096. Section 6 provides a discussion.

# 2 Early-phase vaccine trials

## 2.1 Data collection and structure

We now provide a mathematical description of the HVTN 094 and HVTN 096 data that we study in this paper. The trial schemas and additional scientific details are provided in the Supplementary Material.

The trials randomized healthy, HIV-negative adults to one of several candidate HIV vaccine regimens or placebo. HVTN 094 had three vaccine regimens, labeled T1, T2, and T3, with 10, 15, and 15 participants, respectively. Immune responses were not measured at late time points for T1, so we exclude it from further consideration. HVTN 096 had four vaccine regimens, labeled T1 through T4, each with 20 participants. In both trials, blood specimens were scheduled to be collected during the vaccination series, at two weeks post-last-vaccination—the presumed peak response time point—and at 3 durability time points: 6, 9, and 12 months post-last-vaccination. We let $\mathbf{T}_i = (t_{i1}, \ldots, t_{ik_i})$ be the time (in days) of participant $i$'s observed visits from their presumed peak response visit. These times may vary around the scheduled visit, but are required to fall within pre-specified visit windows. Participants may miss some visits; we will assume that missed visits and visit times are independent of immune responses.

In this article, we measure immune response durability using a summary of the immune responses following the presumed peak time point; any measurements collected before the presumed peak are ignored. This is because vaccines are intended to confer protection once immune responses have reached their peak. Our approach to measuring durability reflects this intention. However, vaccines may confer protection even before the peak response, so it is also of interest to assess durability starting as soon as the first vaccination is administered. We briefly consider extensions to this important problem in the Discussion.

Serum specimens from each visit were used to measure binding antibody responses against multiple HIV-1 antigens – i.e. HIV-1 proteins or portions thereof that elicit an immune response – using the binding antibody multiplex assay (BAMA)[9]. We analyze the outcome $f_{ij}(t) = \log_{10}(R_{ij}(t)/R_{ij0})$, where $R_{ij}(t)$ is participant $i$'s BAMA response for antigen $j$ at $t$ days after the presumed peak time point, and $R_{ij0}$ is

$i$'s baseline immune response. For the BAMA assay, $f_{ij}(t)$ has units $\log_{10}$ net mean fluorescence intensity (MFI). We will refer to the function $f_{ij}(t)$, for $t \geq 0$, as the *post-presumed peak immune response profile* for subject $i$ and antigen $j$. We define $\mathbf{Y}_{ij} = (y_{ij,1}, \ldots, y_{ij,k_i})$ where $y_{ij,k} = f_{ij}(t_{ik})$, and we let participant $i$'s full observed data be denoted by $\mathbf{O}_i = (\mathbf{Y}_{i1}, \ldots, , \mathbf{Y}_{iJ}, \mathbf{T}_i)$. We assume that $\mathbf{O}_1, \ldots, \mathbf{O}_N$ are independent and identically distributed.

## 2.2 Durability parameter of interest

Various summaries of the immune response trajectory have been used to quantify immune response durability. In the immunology literature, classical longitudinal data methods such as mixed effects models have been employed to estimate summaries such as the *half-life* of the immune response (see, e.g.,[3,10–12]). However, the half-life is most meaningful for immune responses that exhibit exponential decay; if the responses do not decay exponentially, then the half-life may not fully describe the kinetics of the decay in immune response, since the rate at which the log-response decays is not constant in time. An alternative measure of the durability of immune responses is the *area under the immune response curve* (AUC) from a pre-defined start time to a pre-defined end time. The area under plasma concentration curves has been used extensively as a summary measure in pharmacokinetics (see, e.g.,[13–15]). However, this literature typically assumes that plasma concentration is measured on a fine time grid. AUC has also been used in the immunology literature to summarize immune response trajectories (see, e.g.,[16–18]). However, to our knowledge there does not yet exist a formal statistical framework for estimating and making inference about the AUC in the setting of early-phase vaccine trials.

Let $\mu_j(t) = E[f_{ij}(t)]$ denote the expected immune response for antigen $j$ at time $t$, where the expectation is taken over a population discussed in detail below. Then $\text{AUC}_j = \int_0^{T_{\max}} \mu_j(t)\, dt$ is the area under the expected immune response curve from the point of presumed peak immune response to $T_{\max}$ days post presumed-peak response (which is fixed across participants). We set $T_{\max}$ to be the number of days between the target presumed-peak response visit and the target day of the final visit, which in the HVTN trials yields $T_{\max} = 351$. Equivalently, by Fubini's Theorem, $\text{AUC}_j = E[\text{AUC}_{ij}]$, where $\text{AUC}_{ij} = \int_0^{T_{\max}} f_{ij}(t)\, dt$. Thus, $\text{AUC}_j/T_{\max}$ is the time- and population-averaged immune response, interpretable as the expected immune response at a uniformly sampled post-presumed-peak time point. The parameter not only has a concrete scientific interpretation, but has relevance for future efficacy testing of the vaccine regimen: if exposure to HIV is assumed to be uniform over zero to $T_{\max}$ days post-peak immune response, $\text{AUC}_j/T_{\max}$ represents the expected immune response at the time of HIV exposure. To compare the durability of immune responses between two vaccine regimens, we use the difference between the respective AUCs. This contrast is denoted

$\Delta\text{AUC}/T_{\max}$, and can be interpreted as the group difference in time- and population-averaged immune responses from presumed peak time point to $T_{\max}$.

In our definition of $\text{AUC}_j$ above, we left the specification of the population arbitrary. We now describe the population of primary interest when profiling the durability of immune response. In assessing durability, our interest is in the sub-population who generate a positive peak immune response, based on an established assay-specific positivity criterion. Subjects with negative responses at the presumed peak time point are expected to have immune responses that fluctuate around the lower limit detectable by the assay, so profiling their decay in response is not of interest. Furthermore, for early-phase trials, our interest lies in understanding the immunological responses generated by the vaccine under ideal conditions. Therefore, we will restrict our attention to the cohort of the observed data who (1) received vaccinations per the study protocol (i.e. received all assigned vaccinations within protocol-specified visit windows), and (2) whose immune responses at the presumed-peak time point were positive according to an assay-specific criterion. Note that condition (2) requires comparing the response at the presumed-peak time point to the immune response at the baseline time point, and has the effect of guaranteeing that the participants included did not miss the peak time point visit. Letting $P_i = 1$ indicate that participant $i$ was observed to receive all assigned vaccinations per-protocol, our estimand of interest is $E[\text{AUC}_{ij} \mid f_{ij}(0) > \tau, P_i = 1]$. Throughout, we make inference conditional on both $f_{ij}(0) > \tau$ and $P_i = 1$, but suppress the conditional notation for brevity.

We note that for early-phase vaccine trials, assessing immune responses at the presumed peak time point, with respect to both positivity and magnitude of response, is typically a primary objective and a first step in evaluating immune response profiles. Any assessment of durability is interpreted in conjunction with an assessment of the peak immune response, so condition (2) would not necessarily have the effect of excluding a vaccine regimen with a high peak immune response but poor durability.

In some cases, our observed-data estimand of interest conditioning on (1) and (2) may have a causal interpretation. Let $f_{ij}^P(t)$ be the potential outcome immune response of participant $i$ to antigen $j$ at $t$ days post-presumed-peak response under per-protocol receipt of the assigned regimen, and let $\text{AUC}_{ij}^P = \int_0^{T_{\max}} f_{ij}^P(t)\,dt$ be the corresponding AUC. Primary interest then lies in the causal estimand $\text{AUC}_j^P = E[\text{AUC}_{ij}^P \mid f_{ij}^P(0) > \tau]$. Under the missing completely at random (MCAR) assumption that $f_{ij}^P$ is independent of $P_i$ (and additional causal assumptions), $\text{AUC}_j^P = E[\text{AUC}_{ij} \mid f_{ij}(0) > \tau, P_i = 1]$. Hence, under MCAR, our observed-data estimand of interest that conditions on (1) and (2) is identified with $\text{AUC}_j^P$.

If MCAR does not hold, then our estimand conditioning on (1) and (2) does not equal $\text{AUC}_j^P$. In these situations, it is common to relax the MCAR assumption to *conditional* independence of $f_{ij}^P$ and $P_i$ given some baseline covariates $W_i$, known as a missing at random (MAR) assumption. Adjustment for baseline covariates may also be necessary for interpretation of the estimand as a causal effect. If participant characteristics

associated with the immune responses are be imbalanced between the conditioned-upon cohorts of the two treatment arms, then our estimand $\Delta\text{AUC}/T_{\max}$ that conditions on (1) and (2) is a net effect, not a causal effect. If the analysis adjusted for all baseline characteristics associated with both presumed-peak response and AUC, the estimand would represent a causal effect. In the Supplementary Material, we comment on how our proposed methods could be modified to adjust for baseline covariates, which could help mitigate violations of MCAR and improve interpretability as a causal effect.

While adjusting for baseline covariates would aid causal interpretability, we also note that in the context of HIV vaccine trials, very little has been found to be predictive of immune responses, despite significant effort. Therefore, it is unlikely that adjusting for observed covariates would drastically alter the results presented here. In trials of candidate vaccines against pathogens such as influenza and dengue viruses, many of the trial participants have had prior exposure to the pathogen. As a result, certain titer concentrations at baseline, such as hemagglutinin antigen titer for influenza or neutralizing antibody titer for dengue, are highly predictive of immune response post-vaccination. However, only participants who test negative to HIV are enrolled in preventive HIV vaccine trials, and so HIV vaccine trial participants typically do not have an immune response at baseline. Furthermore, other available baseline covariates have not been found to be predictive of immune response. For instance,[19] found only weak associations between demographics and neutralizing antibody responses.

While our interest in the analysis of HVTN 094 and HVTN 096 is conditional on requirements (1) and (2), alternative observed-data estimands of interest in other contexts may omit either requirement (1) or (2), or both. The methods defined in the next section can be applied to any of these parameters by using the appropriate cohort for estimation and inference. Requirement (1) may be omitted if interest centers on the intent-to-treat effect rather than the per-protocol effect. Requirement (2) may be omitted if interest centers on the entire cohort, including participants who did not respond to the vaccine.

# 3 Estimators of AUC

In this section we define the four estimators of AUC that we consider. We first focus on defining estimators under MCAR, then briefly discuss extending these estimators under MAR. In defining estimators under MCAR we are implicitly focused on estimation of $\text{AUC}_j$ for a specific antigen $j$ and for a specific vaccine regimen, but we suppress this notation for convenience.

### 3.1 Individually interpolated AUC

We call the first estimator of the AUC the *individually interpolated AUC*. For each participant $i$ with complete post-presumed peak visit data, define $\widehat{\text{AUC}}_{ij}$ as the trapezoid approximation of participant $i$'s individual AUC: $\widehat{\text{AUC}}_{ij} = \sum_{k=2}^{K} \frac{1}{2}(y_{ij,k}+y_{ij,(k-1)})(t_{ik}-t_{i(k-1)})$. The individually interpolated AUC is then the average $\frac{1}{n}\sum_i \widehat{\text{AUC}}_{ij}$ over the $n \leq N$ participants with complete post presumed-peak data. Only participants with complete data are included to ensure that all individual AUCs have the same interpretation.

### 3.2 Interpolated population AUC

The second estimator of $\text{AUC}_j$ we study is the *interpolated population AUC*. Here, for each post-presumed peak visit $k = 1, \ldots, K$, we calculate the average immune response $\bar{y}_{jk}$ and the average follow-up time $\bar{t}_k$ over the participants with non-missing visit $k$ data. The interpolated population AUC is the area under the resulting linearly interpolated population-average immune response curve.

The individually interpolated AUC and interpolated population AUC are attractive estimators because they are simple to compute and model-free. However, both are defined using the trapezoid rule for approximating an integral. Unless the immune responses decline linearly with $t$ between visits, the trapezoid approximation decreases in accuracy as the duration between visits increases. Therefore, we generally expect these AUC estimators to be biased given a finite set of visits. However, as discussed below, we still expect permutation tests based on these estimators to retain the correct type I error rate.

Also, note that the individually interpolated AUC makes use of the individual visit times, but only uses data from fully observed participants. On the other hand, the interpolated population AUC uses all participants' data, but discards the information in the individual visit times; only the average visit times are used. This may constitute a meaningful loss of information for some studies, such as the HVTN studies motivating our work, which use wide visit windows in order to minimize missed visits. The next estimators we consider are based on parametric models for the average immune response curve $\mu_j(t)$, and will allow us to use all the data and the individual visit times.

### 3.3 Linear marginal mean model

The third estimator we consider is based on a linear antibody decay model: $\mu_j(t;\beta_j) = \beta_{0j} + \beta_{1j}t$. Recalling that the outcome of interest $y_{ij}$ is the $\log_{10}$ immune response, the linear model is motivated by the log-linear model commonly fit to antibody decay data (e.g. Amanna et. al., 2007; Yates et. al., 2011; Yates et. al., 2014). The AUC is a simple transformation of $\beta_j$, which we estimate using established Generalized Estimating Equations (GEE) methodology[20].

GEE methods allow for specification of a working within-participant covariance structure, which can sometimes lead to improved estimator efficiency if the working covariance is close to the true covariance. In our data analysis, we use an independence working covariance structure (zeros on the off-diagonal and empirical marginal variances on the diagonal); other structured working covariance structures fit poorly and an unstructured covariance model was unstable given the small sample sizes. We also note that GEE with non-independence working covariance models does not always yield consistent estimators when the conditioned-upon covariates are random [21]. Since visit times are random, this provides additional motivation for using a simple independence working covariance structure in our analysis.

### 3.4 Nonlinear marginal mean model

The fourth estimator we consider is based on a nonlinear antibody decay model. In some cases, the linear model may not fit the data well. For instance, the average immune response might stay elevated before eventually decaying quickly, or decay quickly at first and then level off. While permutation tests for comparing regimens will still have proper type I error rate under model mis-specification, it is possible that a more appropriate marginal mean model will improve power. However, achieving a better fit to the data typically requires more parameters, so there is a trade-off between the potential gain in power due to improved fit and power loss due to estimating additional parameters. We explore this possibility with an estimator based on a nonlinear marginal mean model that we have found to fit the HVTN 094 and HVTN 096 data better than the standard linear model. We consider

$$\mu_j(t;\gamma_j) = \exp\left(\gamma_{0j} + \gamma_{1j} t^{\gamma_{2j}}\right), \tag{1}$$

where $\gamma_{1j}$ is constrained to be negative and $\gamma_{2j}$ is constrained to be positive.

We fit a non-linear GEE model with working independence by minimizing a weighted least-squares criterion. To protect against the possibility that the finite-sample estimating equation has multiple zeros, we also impose a linear constraint on the parameter vector that helps ensure the algorithm finds a sensible zero. A full description of the estimation procedure is provided in the Supplementary Material.

### 3.5 Marginal mean model fits for the HVTN data

Figure 1 illustrates the fit of the above linear and nonlinear models to the HVTN 096 data. Corresponding fits for the HVTN 094 data are provided in the Supplementary Material. While the nonlinear model fits the data quite well for all antigens and vaccine regimens, the linear model appears modestly mis-specified in some instances, e.g. for some antigens in HVTN 096 T1 and T3 (Figure 1). Furthermore, the mis-specification of

the linear model takes a common form wherein the mean at the second time point falls below the linear trend and the mean at the last time point falls above the linear trend. This suggests that the rate of antibody decay (measured in $\log_{10}$ net MFI) may slow down over time. A finer analysis of the rate of antibody decay would require immune response measurements at more time points. The fits of the marginal mean models to the HVTN 094 data revealed a similar pattern for treatment arm T3. It is also apparent from the HVTN 094 fits that certain antigens in HVTN 094 – namely the overall IgG responses to Con S gp140 and gp41 – possess much better durability than any antigens in HVTN 096.

In order to more formally assess the relative fits of the linear and nonlinear models to the data, we compared the leave-one-out mean squared error (MSE) of the linear and nonlinear model fits. For each combination of isotype, antigen, and regimen, we estimated the linear and nonlinear models on each subset of the data formed by leaving one participant's observations out. We then obtained out-of-sample fits of the average response at the observed visit times for the held-out participant, and after repeating this procedure for each participant, computed the average (over all time points and participants) of the squared differences between the fitted means and the true $\log_{10}$ responses.

For 34 out of 46 antigen/regimen combinations, the nonlinear model had strictly smaller leave-one-out MSE than the linear model, and for the remaining 12 combinations it had strictly larger MSE. On average (over the 46 combinations), the nonlinear model had a 5% better MSE than the linear model. The largest improvement of the nonlinear model over the linear model was 19% for the mean vaccine-matched antigens (VMA) of the T1 regimen in HVTN 096. The smallest improvement was -2.7% (the linear model had a smaller MSE than the nonlinear model) for the gp41 antigen of the T3 regimen in HVTN 094.

### 3.6 Uncertainty estimation

As stated previously, we are interested in assessing the null hypothesis that two regimens have the same AUC, or equivalently that $\Delta\text{AUC}/T_{\max} = 0$. To test this null hypothesis we use permutation tests. We pool all the participants in the two regimens, repeatedly randomly permute the pooled participants into two synthetic groups of the same size as the groups in the original data, and, for each permutation, we re-estimate the difference in AUCs between the synthetic groups. The distribution of any statistic over these permutations approximates the distribution of the statistic under the strong null hypothesis that the regimen-specific distributions of the individual-level response curves are identical. Hence the fraction of synthetic differences whose absolute value is larger than the absolute value of the estimated difference in the original data is a valid $p$-value for the test of the null hypothesis that the difference in AUCs is zero.

To construct confidence intervals for $\Delta\text{AUC}/T_{\max}$ between two regimens we use the non-parametric

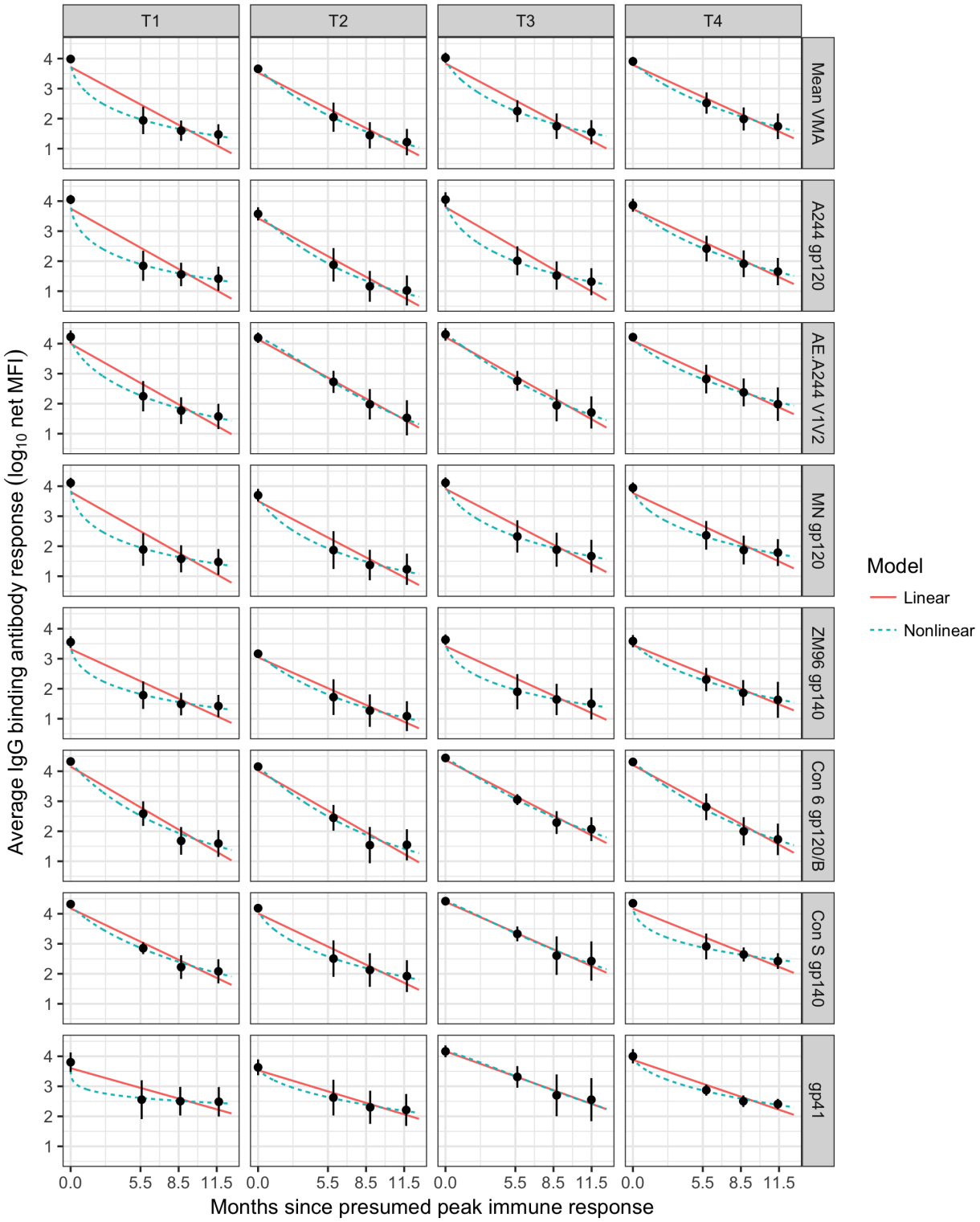


Figure 1: Average post-presumed-peak IgG binding antibody responses and fitted parametric mean models for HVTN 096. Each panel is a particular antigen by vaccine regimen combination. The points are the empirical mean immune responses at the mean follow-up time for each visit and the error bars are 95% confidence intervals for the mean response in each visit. The solid red line is the fitted linear model and the dashed green line is the fitted nonlinear model. "MFI" stands for mean fluorescence intensity.

percentile bootstrap. Here we repeatedly re-sample participants *within* each regimen with replacement and re-estimate $\Delta\text{AUC}/T_{\max}$ between the two bootstrap datasets. Since each participant has a vector of dependent outcomes, this is a cluster bootstrap.

We note that test and confidence intervals based on a sandwich covariance estimator with working independence correlation structure would also yield asymptotically valid tests and confidence intervals. However, in unreported numerical studies, we found that, at the small sample size typical of early-phase vaccine trials, inference based on the sandwich covariance estimator was anti-conservative. We also found this to be the case for certain small-sample corrections to the sandwich estimator (see, e.g. [22]).

# 4 Simulation study

Here, we perform a simulation study to compare the properties of the four estimators defined in Section 3.

## 4.1 Construction

We draw our simulation settings from the motivating data as much as possible. Each simulation setting is based on a pair of two vaccine regimens and antigen-by-isotype combination from the HVTN 094 and HVTN 096 data. We perform simulations both under the null hypothesis that the true difference in AUC is zero and under the alternative hypothesis that it is nonzero. To simulate data under the null, we simulate both data sets from the same vaccine regimen and antigen combination. To simulate data under the alternative, we simulate data sets from each pair of vaccine regimens to be contrasted in HVTN 094 and 096. The nonlinear mean model is used to simulate data under the alternative, as this model was selected for its excellent fit to the data. Under both the null and alternative, for simulated data set we estimate $\Delta\text{AUC}/T_{\max}$ using all four methods described in Section 3 and conduct inference using the permutation test and nonparametric bootstrap. We perform 1000 simulations for each antigen-by-isotype combination.

We simulate data for a given vaccine regimen and antigen-by-isotype combination as follows. First we estimate the parameter $\hat{\gamma}$ of the nonlinear mean model, which will provide our conditional mean model, and we estimate the empirical within-participant correlation matrix of the outcome, which will serve as our within-participant correlation model. Next we compute the empirical mean of the outcome for every visit, and fit a linear model to the log of the variance as a function of the mean, which will serve as our variance model. We then take a bootstrap sample of participant visit data $V$ and $T$. We simulate the outcome for this new design matrix from a multivariate log-normal distribution using the fitted mean of the outcome under the nonlinear model with parameter $\hat{\gamma}$, the empirical correlation matrix, and the fitted variance using the log-linear variance model.

We note that although we simulated data from the nonlinear model, it is not a foregone conclusion that the nonlinear model will outperform the other methods simply because it is correctly specified. The nonlinear model requires estimating one more parameter than the linear model, and given the small sample sizes in this setting, it is possible that estimating this additional parameter is not worth the cost in variability it adds to the estimator.

## 4.2 Results

We now turn to the results of the simulation study. Our overall conclusion is that having a correctly specified marginal mean model is not necessary if the only goal is valid tests comparing vaccine regimens, but it is necessary if the goal is estimation or confidence intervals for $\Delta\text{AUC}/T_{\max}$. If a well-fitting marginal mean model is not available but the decline in immune response is reasonably linear, our simulations suggest that the linear model is preferable to either of the interpolation-based methods because it typically yields a more powerful permutation test.

Figure 2 (top panel) shows the absolute bias of the estimators for $\Delta\text{AUC}/T_{\max}$ under the null and alternative hypotheses. Each point in the figure shows the results for a given simulation setting, i.e. a different pair of vaccine regimens and antigen-by-isotype combination. Under the null hypothesis all estimators have low bias. This is because the data are generated under the null hypothesis that the individual-level immune response profiles follow the same distribution; hence even if an estimator does not recover the true marginal mean function, the estimated mean profiles should be similar and so the estimated $\Delta\text{AUC}/T_{\max}$ is small.

The interpolation-based methods are biased under the alternative hypothesis because the number of time points at which immune responses are measured is small. It is perhaps best to think of the interpolation methods as estimating a different population parameter under the alternative hypothesis. The nonlinear model has the lowest overall bias under the alternative hypothesis because it was used to simulate the data. The linear model has higher average bias than the nonlinear model because it is misspecified.

Figure 2 (bottom panel) shows the coverage of 95% quantile bootstrap confidence intervals for the difference in AUC using the linear and nonlinear models. We omit the interpolation-based methods because, as discussed above, these methods estimate a different parameter under the alternative hypothesis, and hence cannot be used to give accurate confidence intervals. Under the null hypothesis the bootstrap is generally slightly anti-conservative. Under the alternative hypothesis, the linear model is severely anti-conservative due to model mis-specification, and the correctly-specified nonlinear model provides improved coverage.

Figure 3 (top panel) shows the size of the permutation test using the four estimators. The permutation test has valid size for each of the four estimators. Hence, as discussed in Section 3, any of the four estimators

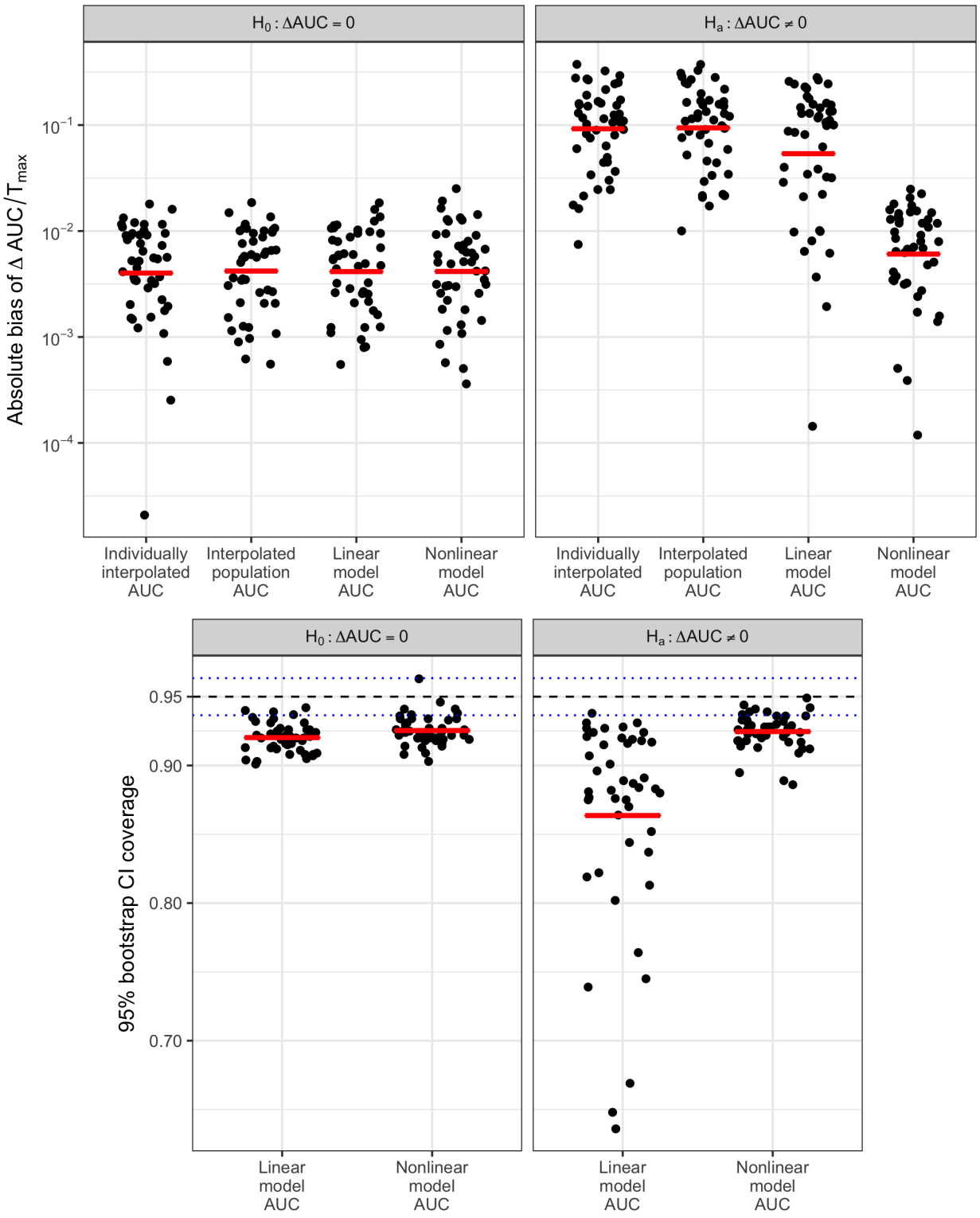


Figure 2: Top: Absolute bias of the four estimators of $\Delta\text{AUC}/T_{\max}$. Bottom: Coverage of 95% bootstrap CIs for the true $\Delta\text{AUC}/T_{\max}$ using the linear and nonlinear models. The different points correspond to different simulation settings, all of which were drawn from the HVTN 094 and 096 data. The left panels are under the null hypothesis (the true difference is zero) and the right panels are under various alternative hypotheses (non-zero differences). The red solid lines are the average absolute biases over the simulation settings. The black dashed line is the nominal 95% coverage and the blue dotted lines represent Monte Carlo error.

can be used in conjunction with a permutation test as a valid means of testing the strong null hypothesis that two regimens have the same post-presumed-peak marginal mean immune response profile.

Figure 3 (bottom panel) shows the power of the permutation test using the four estimators. Overall, the nonlinear model has the largest power of the four estimators because it is correctly specified. Of the three other estimators, the linear model has slightly larger average power than the interpolation-based methods. We also see that there are some alternative hypotheses for which all four estimators have very low power—when alternatives are very close to the null—and a few alternatives for which all four estimators have very high power—when alternatives are very far from the null.

Figure 4 plots the difference in power between the first three estimators and the correctly-specified nonlinear model against the absolute value of the true difference in AUC. In general, the difference in power correlates with the true effect size: the increase in power for the correctly-specified nonlinear model is largest when the true effect size is large. When the true effect size is small, the extra variability produced by fitting the additional parameter in the nonlinear mean model sometimes yields worse power than the other methods. The largest difference in power between the linear model and the nonlinear model was near 0.2, but when one of the first three estimators had greater power most of the power differences were under 0.1. The other major source of variability in the differences in power is the extent of the nonlinearity of the underlying data. When the true average immune response curve is close to linear, the nonlinear model is unnecessary.

We also conducted a simulation study to assess the performance of the four estimators under simulation from the linear model. The setup of this study was identical to that described above, except that the linear model was used as the true marginal mean model rather than the nonlinear model. The true parameter $\beta_0$ of the linear mean model used for simulation was obtained by estimating the linear mean model for each given vaccine regimen and antigen-by-isotype combination.

The results of this second study are presented in the Supplementary Material. Our overall conclusion is that, under model mis-specification, the estimator based on the nonlinear model generally retains valid inferential properties, but can suffer from a serious loss of power compared to use of the other three estimators in the context of small sample sizes. This is because the nonlinear mean model requires estimating an additional parameter, which adds to the variance in the estimator. Unsurprisingly, the estimator based on the linear model performs best under simulation from the linear model. Additionally, we conclude that the interpolation-based estimators perform substantially better under simulation from the linear model than under simulation from the nonlinear model. This is also unsurprising, since these methods employ a linear interpolation between time points, and thus would be expected to perform well if the truth is linear. These results highlight the need for caution when adding even a single additional parameter to a model in the context of small$-n$ studies: unless the more flexible model clearly fits better than a simpler model, in many

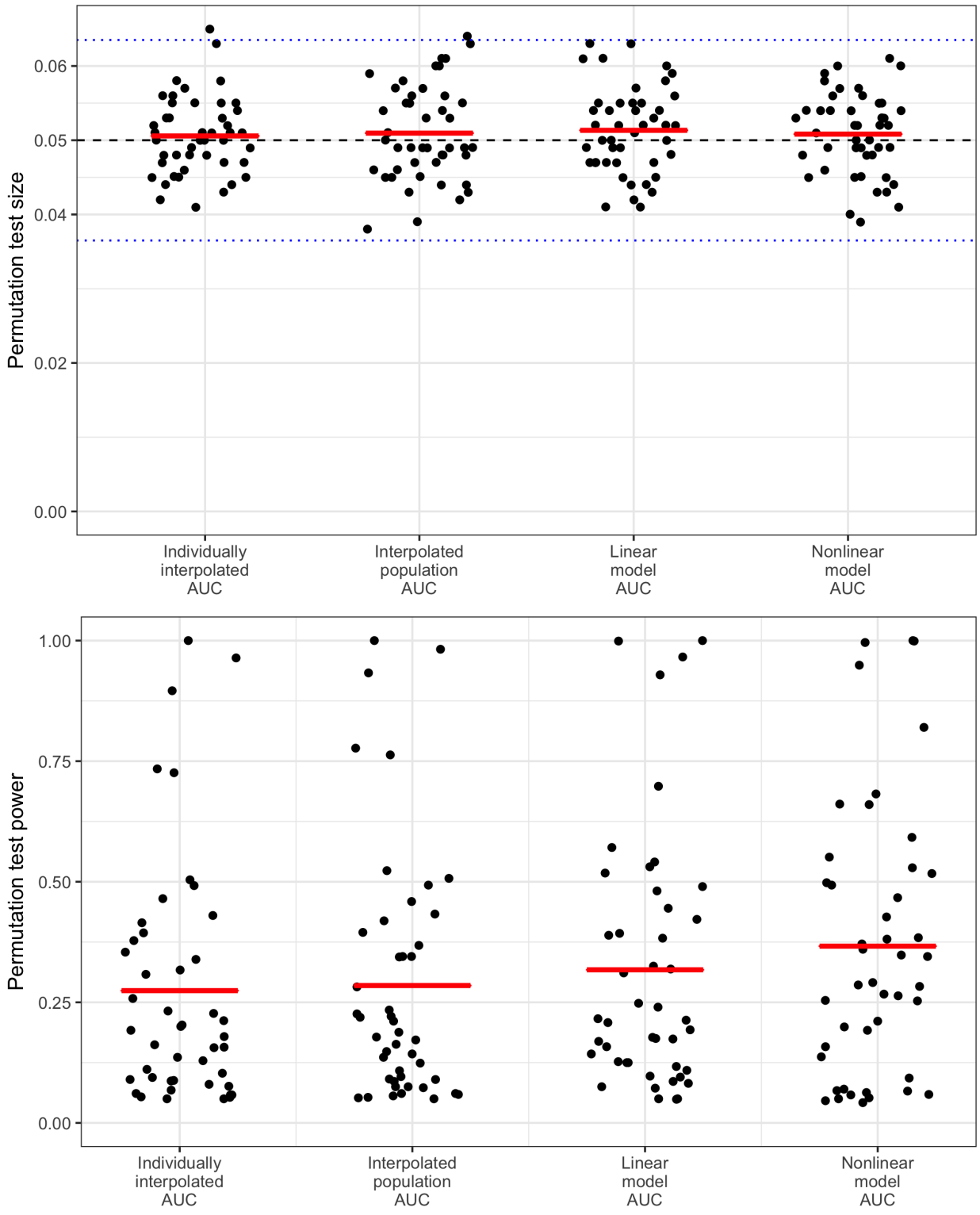


Figure 3: Top panel: Size of the permutation test for the four methods. Bottom panel: Power of the permutation test for the four methods. In both panels, the different points correspond to different simulation settings, all of which were drawn from the real data. Each point is the proportion of permutation test $p$-values less than 0.05 out of 1000 simulations. In the top panel, the black dashed line is the nominal 0.05 level and the blue dotted lines represent Monte Carlo error. The red solid lines are the average sizes and powers over the simulation settings.

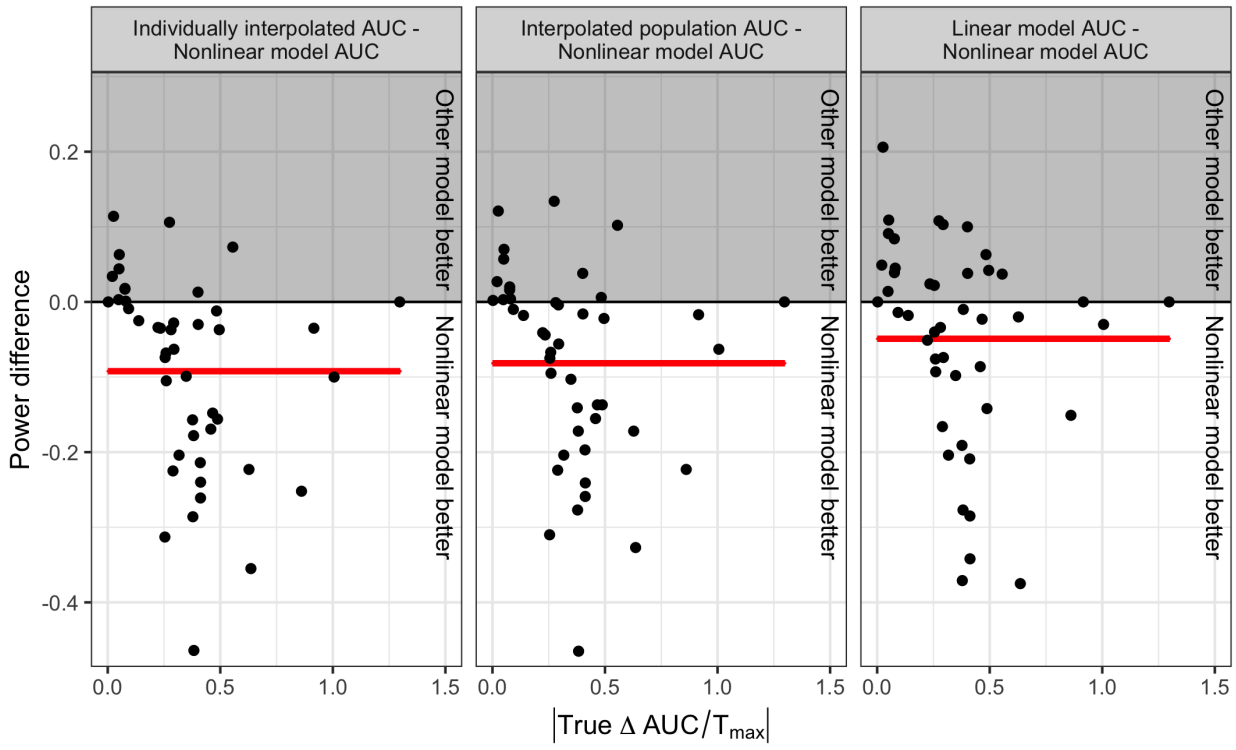


Figure 4: Difference in power of inferential methods, compared to the nonlinear model, as a function of the true $\Delta\text{AUC}/T_{\max}$. The $y$-axis is the difference in power of the permutation test between each method and the nonlinear model. The $x$-axis is the absolute value of the true $\Delta\text{AUC}/T_{\max}$. Each point represents the proportion of 1000 simulations rejecting the null hypothesis of no difference at level 0.05. The red solid lines are the average power differences over the simulation settings.

cases it may not be worth the cost in power.

# 5 Comparison of vaccine regimens within and between HVTN 094 and HVTN 096

In this section we compare the durability of the humoral responses elicited by the HVTN 094 and HVTN 096 vaccine regimens. Scientific details of the vaccine regimens and antigens of interest in the two studies are provided in the Supplementary Material.

For all analyses we use the proposed nonlinear marginal mean model to estimate $\Delta\text{AUC}/T_{\max}$, the nonparametric bootstrap to obtain confidence intervals for $\Delta\text{AUC}/T_{\max}$, and permutation tests to obtain $p$-values of the null hypothesis that $\Delta\text{AUC}/T_{\max} = 0$. Permutation test $p$-values greater than or less than 0.1 are based on $10^3$ or $10^4$ random permutations, respectively. We also include $q$-values that adjust for multiple antigen tests within each vaccine comparison using the false discovery rate method [23]. Throughout,

| Isotype | Antigen | $\Delta\text{AUC}/T_{\max}$ | 95% CI | $p$-value | $q$-value |
|---|---|---|---|---|---|
| IgG | **Con 6 gp120/B** | **0.86** | **[ 0.053, 1.7 ]** | **0.050** | **0.075** |
| | **Con S gp140** | **0.38** | **[ 0.070, 0.74 ]** | **0.032** | **0.075** |
| | **gp41** | **0.25** | **[ -0.019, 0.55 ]** | **0.14** | **0.14** |
| IgG1 | Con S gp140 | 0.47 | [ -0.059, 0.95 ] | 0.11 | 0.22 |
| | gp41 | 0.29 | [ -0.25, 0.87 ] | 0.28 | 0.28 |
| IgG3 | Con S gp140 | 0.46 | [ -0.13, 0.96 ] | 0.14 | 0.27 |
| | gp41 | 0.32 | [ -0.37, 0.93 ] | 0.40 | 0.40 |

Table 1: Comparison of the durability of the humoral responses elicited by T2 versus T3 within HVTN 094. Positive values indicate that T2 elicited a larger average humoral response than T3. $\Delta\text{AUC}/T_{\max}$ is the difference of average humoral responses with units $\log_{10}$ net MFI. $p$-values are two-sided and $q$-values adjust for multiple antigen tests using the method of[23] within a given IgG isotype. Comparisons with $q$-values less than 0.2 are shown in boldface.

$q$-values less than 0.2 are considered to indicate statistical significance.

Figure 1 (and Figure 8 in the Supplementary Material) show the estimated immune response curves, along with empirical visit means and the estimated nonlinear model, for the HVTN 094 (and HVTN 096) data. Tables 1 and 2 (and Figures 10 and 11 in the Supplementary Material) present the $\text{AUC}/T_{\max}$ estimates and confidence intervals for each isotype-antigen-regimen combination in HVTN 094 (and HVTN 096).

In HVTN 094 we compare vaccine regimens T2 and T3, which differ in that T2 included an extra dose of the MVA62B vaccine component as compared to T3, to assess whether the extra MVA62B dose had a significant effect on durability. Table 1 shows the results of these comparisons. We find that T2 elicited a moderately larger average IgG response than T3, suggesting that the extra dose of MVA62B improved immune response durability as measured by the AUC.

In HVTN 096 we compare T1 vs. T2 and T3 vs. T4. These regimens differ in that T2 and T4 regimens co-administered AIDSVAX® protein during the priming, at months 0 and 1, and regimens T1 and T3 did not. We also compare T1 vs. T3 and T2 vs. T4, regimens which differ in the prime– NYVAC or plasmid DNA– with and without AIDSVAX® protein co-administration. Table 2 shows the results of these comparisons. We find that T3 elicited immune responses with greater durability than T1, and that T4 durability was greater than T2, indicating that DNA priming yielded more durable immune responses than NYVAC priming, both with and without co-administration of AIDSVAX® at the prime. However, we find the magnitude of the estimated increase in durability to be small relative to the overall level of the immune responses. We did not find a significant benefit of co-administration of AIDSVAX® during the prime.

Finally, we compare HVTN 094 T2 to HVTN 096 T3 and T4. Each regimen used the DNA as the prime at months 0 and 1; however, in HVTN 094 the boost was MVA62B, while for HVTN 096 the boost was NYVAC plus AIDSVAX®. Table 3 shows the results of these comparisons. We find that HVTN 094 T2 induced immune responses with greater AUC than both HVTN 096 T3 and T4 for Con S gp140 and gp41,

| Comparison | Antigen | $\Delta$AUC/$T_{\max}$ | 95% CI | $p$-value | $q$-value |
|---|---|---|---|---|---|
| T1 - T2 | Mean VMA | -0.08 | [ -0.54 , 0.37 ] | 0.74 | 0.94 |
| | A244 gp120 | 0.026 | [ -0.53, 0.61 ] | 0.92 | 0.94 |
| | AE.A244 V1V2 | -0.41 | [ -0.93, 0.03 ] | 0.11 | 0.90 |
| | MN gp120 | 0.05 | [ -0.52, 0.65 ] | 0.86 | 0.94 |
| | ZM96 gp140 | 0.051 | [ -0.47, 0.61 ] | 0.85 | 0.94 |
| | Con 6 gp120/B | 0.048 | [ -0.38, 0.54 ] | 0.83 | 0.94 |
| | Con S gp140 | 0.26 | [ -0.18, 0.78 ] | 0.29 | 0.94 |
| | gp41 | -0.02 | [ -0.68, 0.6 ] | 0.94 | 0.94 |
| T3 - T4 | Mean VMA | -0.22 | [ -0.64, 0.19 ] | 0.34 | 0.45 |
| | A244 gp120 | -0.38 | [ -0.95, 0.13 ] | 0.19 | 0.38 |
| | AE.A244 V1V2 | -0.077 | [ -0.53, 0.37 ] | 0.75 | 0.85 |
| | MN gp120 | -0.0024 | [ -0.66, 0.61 ] | 1.0 | 1.0 |
| | ZM96 gp140 | -0.29 | [ -0.91, 0.21 ] | 0.33 | 0.45 |
| | Con 6 gp120/B | 0.28 | [ -0.076, 0.66 ] | 0.17 | 0.38 |
| | Con S gp140 | 0.38 | [ -0.029, 0.76 ] | 0.064 | 0.26 |
| | gp41 | 0.38 | [ -0.019, 0.75 ] | 0.05 | 0.26 |
| T1 - T3 | Mean VMA | -0.26 | [ -0.71, 0.21 ] | 0.26 | 0.36 |
| | A244 gp120 | -0.093 | [ -0.68, 0.49 ] | 0.74 | 0.74 |
| | AE.A244 V1V2 | -0.41 | [ -0.94, 0.053 ] | 0.11 | 0.22 |
| | MN gp120 | -0.35 | [ -0.95, 0.28 ] | 0.27 | 0.36 |
| | ZM96 gp140 | -0.14 | [ -0.71, 0.45 ] | 0.63 | 0.72 |
| | **Con 6 gp120/B** | **-0.49** | **[ -0.82, -0.16 ]** | **0.0087** | **0.035** |
| | **Con S gp140** | **-0.41** | **[ -0.70, -0.13 ]** | **0.0069** | **0.035** |
| | **gp41** | **-0.64** | **[ -1.3, -0.066 ]** | **0.033** | **0.089** |
| T2 - T4 | **Mean VMA** | **-0.40** | **[ -0.81, -0.0011 ]** | **0.073** | **0.20** |
| | **A244 gp120** | **-0.50** | **[ -1.0, -0.013 ]** | **0.062** | **0.20** |
| | AE.A244 V1V2 | -0.076 | [ -0.51, 0.38 ] | 0.80 | 0.80 |
| | MN gp120 | -0.40 | [ -0.96, 0.22 ] | 0.21 | 0.42 |
| | **ZM96 gp140** | **-0.48** | **[ -1.0, -0.026 ]** | **0.073** | **0.20** |
| | Con 6 gp120/B | -0.25 | [ -0.74, 0.20 ] | 0.33 | 0.44 |
| | Con S gp140 | -0.29 | [ -0.90, 0.26 ] | 0.32 | 0.44 |
| | gp41 | -0.23 | [ -0.73, 0.21 ] | 0.39 | 0.45 |

Table 2: Comparison of the durability of the humoral immune responses elicited by the different vaccine regimens within HVTN 096. Positive values indicate that the vaccine regimen on the left elicited immune responses with greater durability than those elicited by the regimen on the right. VMA stands for vaccine-matched antigens, which for HVTN 096 were A244 gp120, AE.A244 V1V2, MN gp120, and ZM96 gp140. $\Delta$AUC/$T_{\max}$ is the average immune response with units $\log_{10}$ net MFI. $p$-values are two-sided and $q$-values adjust for multiple antigen tests within each regimen comparison using the method of[23]. Comparisons with $q$-values less than 0.2 are shown in boldface.

but had a significantly smaller AUC than HVTN 096 T3 for Con 6 gp120/B. The sizes of these effects are moderate relative to the average immune responses. The estimated improvement in IgG response durability with MVA62B boosting as opposed to NYVAC plus AIDSVAX® boosting was greater when AIDSVAX® was given at the prime. Figure 9 in the Supplementary Material shows the nonlinear-model-based and empirical mean immune response profiles for these cross-protocol comparisons.

| Comparison | Antigen | $\Delta\text{AUC}/T_{\max}$ | 95% CI | $p$-value | $q$-value |
|---|---|---|---|---|---|
| 096 T3 - 094 T2 | **Con 6** | **0.56** | **[ 0.096, 0.97 ]** | **0.0075** | **0.0075** |
| | **Con S** | **-0.63** | **[ -0.92, -0.33 ]** | **0.0013** | **0.0020** |
| | **gp41** | **-0.92** | **[ -1.3, -0.55 ]** | $\mathbf{3.0 \times 10^{-4}}$ | $\mathbf{9.0 \times 10^{-4}}$ |
| 096 T4 - 094 T2 | Con 6 | 0.27 | [ -0.30, 0.78 ] | 0.42 | 0.42 |
| | **Con S** | **-1.0** | **[ -1.4, -0.61 ]** | $\mathbf{1.0 \times 10^{-4}}$ | $\mathbf{1.0 \times 10^{-4}}$ |
| | **gp41** | **-1.3** | **[ -1.5, -1.1 ]** | $\mathbf{1.0 \times 10^{-4}}$ | $\mathbf{1.0 \times 10^{-4}}$ |

Table 3: Comparison of the durability of humoral immune responses elicited by vaccine regimens in HVTN 094 versus HVTN 096. Positive values indicate that the HVTN 096 vaccine regimen elicited immune responses with better durability than those elicited by HVTN 094 T2. $\Delta\text{AUC}/T_{\max}$ is the average immune response with units $\log_{10}$ net MFI. $p$-values are two-sided and $q$-values adjust for multiple antigen test within each regimen comparison using the method of[23]. Comparisons with $q$-values less than 0.2 are in boldface. Con 6 is short for Con 6 gp120/B, and Con S is short for Con S gp140.

# 6 Discussion

Using the difference in the areas under the marginal mean log antibody response curves, we compared the durability of antibody responses elicited by various vaccine regimens tested in early-phase HIV vaccine trials. We compared four estimators for this parameter, three of which have been employed in the literature, and one of which uses a nonlinear marginal mean model and is thus tailored to data from two phase I HIV vaccine trials, HVTN 094 and HVTN 096. Our simulation study indicated that any of the methods can be used to obtain valid permutation-based hypothesis tests, but a well-fitting mean model is necessary for precise estimation of area under the curve, and to obtain good bootstrap confidence interval coverage. A correctly-specified mean model also generally increases power for detecting differences in durability between vaccine regimens, but only if the effect is large enough to outweigh additional free parameters in the model. When a simpler model is only modestly mis-specified, or not at all, additional free parameters can come with a significant cost to power, which suggests that substantial caution should be used when adding parameters to the model in this context.

We used the nonlinear marginal mean model to compare the durability of immune responses between vaccine regimens within and between HVTN 094 and HVTN 096. In HVTN 096, we found that priming with plasmid DNA as opposed to NYVAC, with or without co-administration of AIDSVAX® protein, generated more durable immune responses to certain HIV antigens. We also found that the extra dose of MVA62B in

HVTN 094 had a significant, albeit small, positive effect on humoral response durability. Comparing across protocols, we found that the MVA62B boost in HVTN 094 yielded greater humoral response durability than the NYVAC plus AIDSVAX® boost in HVTN 096. If higher IgG and IgG subclass immune responses to the antigens tested translate into greater protection against HIV, these results suggest that the MVA62B vaccine has the potential to produce longer-lasting protection. It would be of interest to use the methods presented here to assess and compare durability of other immune responses induced by these regimens, e.g. cellular immune responses and antibody-dependent cell-mediated cytotoxicity (ADCC) responses. It would also be of interest to use the methods presented here to assess durability in the intent-to-treat cohort, rather than restricting to participants who received vaccinations per-protocol.

In practice, we suggest using the approach employed here to find a well-fitting marginal mean model. Assessing model fit is relatively simple using visual inspection of the model-predicted vs. observed marginal means and leave-one-out cross validation to estimate mean squared errors. We arrived at our particular nonlinear model by considering simple parametric families on $[0, \infty)$ that are bounded, positive, smooth, monotonically decreasing, and permit varying degrees of nonlinear decay. Given the relatively short one-year follow-up we did not anticipate observing more than one phase of antibody decay [1]. However, the nonlinear model we have employed has some drawbacks. First, while it fits the IgG immune response data from the HVTN trials well, it may not fit assay data of other types. Second, it has an asymptote at zero and hence implies complete decay of the marginal mean log response eventually. Third, the linear and nonlinear models are not nested, so there is no single parameter that can be used to test sufficiency of the linear model. However, as Figure 1 demonstrates, the nonlinear model can approximate a linear model very well.

It is of interest to further investigate optimal methods for summarizing not just post-peak but full immune response profiles. We note that both the linear and nonlinear mean models we explored are monotonic, and hence cannot be used to model pre-presumed peak data. In contrast, the trapezoid-rule methods are comparatively attractive because they can be used for pre-presumed peak data and do not imply a functional form of the response. However, they estimate a parameter that depends on the time points at which the immune responses are measured, which is an unattractive attribute when comparing across trials with different visit schedules. It would be desirable to have statistical methods for this setting that avoid overly restrictive parametric assumptions but that do not estimate data-dependent parameters or discard data as do the trapezoid-rule methods. In applying the methods studied here to vaccines against seasonal pathogens such as influenza or malaria, it would also be of interest to incorporate a weight function over the follow-up period. These are topics for future research.

While we have focused here on population-average quantities as summaries of durability, an alternative approach is to measure durability using a summary of the average individual in the population. This is

the approach taken, for instance, by mixed effects models. However, mixed-effects models often rely on parametric assumptions regarding the distribution of individual-level random effects, whose validity can be difficult to assess. Methods for estimating population-average quantities, such as those we have presented here, do not require specifying or assessing such distributional assumptions.

The antibody half-life has been used elsewhere to measure durability of vaccine-induced immune responses[3,10,11]. However, the utility of this metric is limited to settings where the individual immune response curves decay exponentially; when these curves exhibit non-exponential decay, as in our motivating data, the interpretation of the half-life is less straightforward. We also note that in principle we could use our non-linear marginal mean model to compute the half-life of the average immune response, and compare regimens in terms of their half-lives.

[24] recently assessed the ability of various types of HIV vaccine-induced immune responses measured two weeks post-last vaccination to predict immune responses measured six months later, for settings where resources are spared by not measuring those later immune responses. The goals of the present article were different in that we aimed to compare the durability of various vaccine regimens using measured post-presumed peak immune responses, as opposed to predicted responses. This type of analysis can help guide the selection of vaccine regimens for future trials.

The permutation test we employed for comparing the durability of immune responses between vaccine regimens does not apply directly to comparisons of immune response durability between antigens within vaccine regimen, since responses to all antigens are measured on the same participants and therefore the data are dependent. Extending the testing methods presented here to such comparisons would be an interesting topic for future work.

# Acknowledgements

This work was supported by the National Institute of Allergy and Infectious Disease at the National Institutes of Health (UM1 AI068618 to J. M., UM1 AI068635 to P. B. G., and UM1 AI068614 to L. C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We would like to thank the parent protocol teams, site staff, and trial participants of HVTN 094 (ClinicalTrials.gov identifier NCT01571960) and HVTN 096 (ClinicalTrials.gov identifier NCT01799954). We would also like to thank Lindsay Carpp for help with manuscript editing.

# Supplementary Material

Supplementary Material for this paper presents the HVTN 094 and HVTN 096 schemas, details for the GEE procedure used to fit the nonlinear mean model, details for estimation under MAR, results of the simulation study under the linear model, spaghetti plots of the HVTN 094 and HVTN 096 immune response data, model fits for the HVTN 094 data, and estimates and confidence intervals for $AUC/T_{\max}$ from the HVTN 094 and HVTN 096 regimens.

# References

1. Lewis G, DeVico A and Gallo R. Antibody persistence and T-cell balance: Two key factors confronting HIV vaccine development. *Proceedings of the National Academy of Sciences* 2014; 111: 15614–15621.

2. Robb M, Rerks-Ngarm S, Nitayaphan S et al. Risk behaviour and time as covariates for efficacy of the HIV vaccine regimen ALVAC-HIV (vCP1521) and AIDSVAX B/E: a post-hoc analysis of the Thai phase 3 efficacy trial RV 144. *The Lancet Infectious Diseases* 2012; 12: 531–537.

3. Yates N, Liao H, Fong Y et al. Vaccine-induced Env V1-V2 IgG3 correlates with lower HIV-1 infection risk and declines soon after vaccination. *Science Translational Medicine* 2014; 6.

4. DeVico AL, Lewis GK and Gallo RC. Modulating the durability of anti-HIV gp120 antibody responses after vaccination: a comment on Wilson & Karp. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2015; 370: 20150199.

5. Yates NL, deCamp AC, Korber BT et al. HIV-1 Envelope Glycoproteins from Diverse Clades Differentiate Antibody Responses and Durability among Vaccinees. *Journal of Virology* 2018; .

6. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine* 2009; 361: 2209–2220.

7. Haynes B, Gilbert P, McElrath M et al. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *New England Journal of Medicine* 2012; 366: 1275–1286.

8. Pantaleo G, Janes H, Tomaras G et al. Comparing Different Priming Strategies to Optimize HIV Vaccine Antibody Responses: Results from HVTN 096/EV04 (NCT01799954). *AIDS Research and Human Retroviruses* 2016; 32: 68–68.

9. Tomaras G, Yates N, Liu P et al. Initial B-cell responses to transmitted human immunodefiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *Journal of Virology* 2008; 82: 12449–12463.

10. Amanna I, Carlson N and Slifka M. Duration of humoral immunity to common viral and vaccine antigens. *New England Journal of Medicine* 2007; 357: 1903–1915.

11. Yates N, Lucas J, Nolen T et al. Multiple HIV-1-specific IgG3 responses decline during acute HIV-1: implications for detection of incident HIV infection, 2011.

12. Francica JR, Zak DE, Linde C et al. Innate transcriptional effects by adjuvants on the magnitude, quality, and durability of HIV envelope responses in NHPs. *Blood Advances* 2017; 1: 2329–2342.

13. Yamaoka K, Nakagawa T and Uno T. Statistical moments in pharmacokinetics. *J Pharmacokin Biopharm* 1978; 6: 547–558.

14. Bourne D. *Mathematical Modeling of Pharmacokinetic Data*. New York: Routledge, 1995.

15. Jawień W. Searching for an optimal AUC estimation method: a never-ending task? *Journal of Pharmacokinetics and Pharmacodynamics* 2014; 41(6): 655–673. DOI:10.1007/s10928-014-9392-y. URL `https://doi.org/10.1007/s10928-014-9392-y`.

16. Morse M, Hobeika A, Osada T et al. Depletion of human regulatory T cells specifically enhances antigen-specific immune responses to cancer vaccines. *Blood* 2008; 112: 610–618.

17. R T, K K, E M et al. Serum indoleamine 2,3-dioxygenase activity is associated with reduced immunogenicity following vaccination with MVA85A. *BMC Infectious Diseases* 2014; 14.

18. Palgen J, Tchitchek N, Elhmouzi-Younes J et al. Prime and boost vaccination elicit a distinct innate myeloid cell immune response. *Scientific Reports* 2018; 8: 3087.

19. Gilbert P, Wang M, Wrin T et al. Magnitude and Breadth of a Nonprotective Neutralizing Antibody Response in an Efficacy Trial of a Candidate HIV-1 gp120 Vaccine. *The Journal of Infectious Diseases* 2010; 202(4): 595–605.

20. Zeger S and Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; 42: 121–130.

21. Pepe MS and Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics – Simulation and Computation* 1994; 23(4): 939–951.

22. Fay MP and Graubard BI. Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. *Biometrics* 2001; 57(4): 1198–1206.

23. Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 1995; 57: 289–300.

24. Huang Y, Zhang L, Janes H et al. Predictors of durable immune responses six months after the last vaccination in preventive HIV vaccine trials. *Vaccine* 2017; 35: 1184–1193.