Research Article

Zhe Chen, Xinran Li and Bo Zhang*

The role of randomization inference in unraveling individual treatment effects in early phase vaccine trials

https://doi.org/10.1515/scid-2024-0001
Received February 25, 2024; accepted July 19, 2024; published online August 12, 2024

Abstract: Randomization inference is a powerful tool in early phase vaccine trials when estimating the causal effect of a regimen against a placebo or another regimen. Randomization-based inference often focuses on testing either Fisher's sharp null hypothesis of no treatment effect for any participant or Neyman's weak null hypothesis of no sample average treatment effect. Many recent efforts have explored conducting exact randomization-based inference for other summaries of the treatment effect profile, for instance, quantiles of the treatment effect distribution function. In this article, we systematically review methods that conduct exact, randomization-based inference for quantiles of individual treatment effects (ITEs) and extend some results to a special case where naïve participants are expected not to exhibit responses to highly specific endpoints. These methods are suitable for completely randomized trials, stratified completely randomized trials, and a matched study comparing two non-randomized arms from possibly different trials. We evaluate the usefulness of these methods using synthetic data in simulation studies. Finally, we apply these methods to HIV Vaccine Trials Network Study 086 (HVTN 086) and HVTN 205 and showcase a wide range of application scenarios of the methods. R code that replicates all analyses in this article can be found in first author's GitHub page at https://github.com/Zhe-Chen-1999/ITE-Inference.

Keywords: causal inference; early phase clinical trials; immunogenicity; effect quantile; randomization inference; vaccine

Introduction

Early phase vaccine trials; vaccine-induced immune responses; heterogeneity

One primary objective stated in study protocols of early phase clinical trials of experimental vaccines is to evaluate the vaccine-induced immunogenicity. Vaccine-induced immune responses are often heterogeneous among study participants. To illustrate, Figure 1 exhibits the observed serum IgG binding antibody multiplex assay (BAMA) responses to two antigens, Con 6 gp120/B and gp41, among study participants in a phase 1, multi-arm, placebo-controlled clinical trial conducted via the HIV Vaccine Trials Network (HVTN) [1, 2]. HVTN 086/SAAVI 103 (HVTN 086 henceforth) enrolled a total of 184 participants into four study arms; within each study arm,

***Corresponding author: Bo Zhang**, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA, E-mail: bzhang3@fredhutch.org. https://orcid.org/0000-0002-4381-1124
Zhe Chen, Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL, USA, E-mail: zhec6@illinois.edu
Xinran Li, Department of Statistics, University of Chicago, Chicago, IL, USA, E-mail: xinranli@uchicago.edu

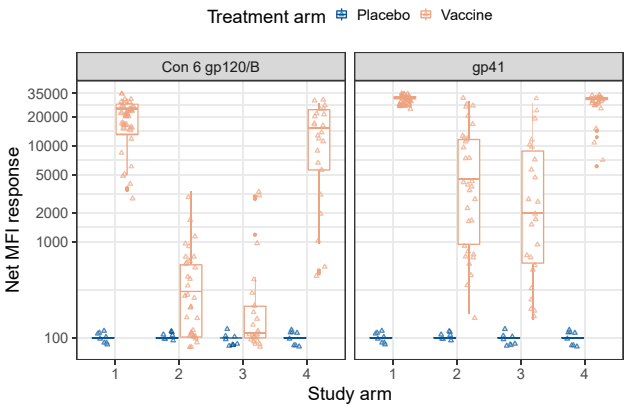


Figure 1: Observed serum IgG BAMA responses to Con6 gp120/B (left panel) and gp41 (right panel) among study participants in the HIV Vaccine Trials Network (HVTN) study 086. A total of four study arms were plotted. Within each arm, participants were randomized to a vaccine regimen or placebo. A small perturbation was added to each observation to aid data visualization.

participants were randomized to a candidate vaccine regimen or placebo. It is transparent from Figure 1 that vaccine recipients' binding antibody responses ranged from "potent and mostly homogeneous" (e.g., response to gp41 among recipients of regimens 1 and 4) to "weak and heterogeneous" (e.g., response to Con 6 gp120/B among recipients of regimen 2). This within-participant heterogeneity in vaccine-induced immune responses has been well-documented in many vaccines, including those for Covid-19, influenza, dengue, and hepatitis B, [2] and could at least partially explain the lack of efficacy in phase 2b/3, HIV-1 vaccine trials.

Researchers routinely characterize a vaccine regimen's induced immune responses (e.g., the BAMA responses in Figure 1) by estimating and reporting its sample average treatment effect (SATE) against the placebo, which could be unbiasedly estimated in a randomized clinical trial, and assess and rank multiple regimens by comparing their estimated SATEs. When developing a challenging vaccine product like an HIV-1 vaccine, researchers have long realized that the response rates among study participants are often highly variable, and a significant proportion of participants could exhibit immune responses below the assay limit of detection (LOD) or lower limit of quantification (LLOQ). Hence, summarizing and comparing immune response profiles based on the mean difference alone could mask significant and perhaps meaningful heterogeneities. The current standard practice is to complement the estimated SATE and its 95 % confidence interval by further reporting (i) descriptive statistics and boxplots summarizing the spread of immune responses elicited by each regimen, (ii) the percentage of positive or high responders within each group, and (iii) the mean response among the subset of positive or high responders. Unfortunately, unlike the SATE which is a well-defined, albeit less than comprehensive, causal estimand, neither the descriptive statistics nor the mean difference in responses among positive or high responders constitutes a formal confidence statement about the "treatment effect" of a regimen vs. placebo or another regimen.

Science table; estimands of interest; outline of the article

What is a well-defined causal estimand that captures treatment effect heterogeneity in early phase clinical trials? It is instructive to examine what Rubin [3] refers to as a "science table." Table 1 summarizes the potential outcomes and unit-level treatment effects of $N_1 + N_0$ study participants in a clinical trial, where N_1 study participants are randomized to Regimen 1 and the other N_0 participants to Regimen 0. In the science table, $Y_i(1)$ and $Y_i(0)$ denote study participant i's potential immune responses of interest under two regimens, respectively, and only one of the two responses is observed depending on the actual regimen assigned to the participant. The contrast in the two potential outcomes, $\tau_i = Y_i(1) - Y_i(0)$, denotes the unit-level treatment effect [3]. We will refer to τ_i as an *individual treatment effect* (ITE) in this article [4, 5]. The collection of ITEs, $\mathcal{T} = \{\tau_i,\ i = 1, \ldots, N_1 + N_0\}$, is the causal quantity of *ultimate interest*, in the sense that any summary treatment effect can be derived from $\mathcal{T}$ (e.g., the sample average treatment effect $\overline{\tau}$). Let $\tau_{(i)}$ denote the ith largest treatment effect. The immunogenicity profile of a vaccine regimen (against placebo or a competing regimen), as revealed in an early phase

Table 1: Science table of $N_0 + N_1$ study participants. A total of N_0 are randomized to regimen 0 and the other N_1 are randomized to regimen 1. Each participant is associated with two potential immune responses $Y_i(0)$ and $Y_i(1)$ corresponding to regimen 0 and 1, though only the one corresponding to the actual regimen assignment is observed (**boldface**). Each participant i is associated with an individual treatment effect τ_i.

Participants	Regimen	$Y(1)$	$Y(0)$	Individual treatment effects
1	Regimen 1	$\mathbf{Y_1(1)}$	$Y_1(0)$	$\tau_1 = Y_1(1) - Y_1(0)$
2	Regimen 1	$\mathbf{Y_2(1)}$	$Y_2(0)$	$\tau_2 = Y_2(1) - Y_2(0)$
⋮	⋮	⋮	⋮	⋮
N_1	Regimen 1	$\mathbf{Y_{N_1}(1)}$	$Y_{N_1}(0)$	$\tau_{N_1} = Y_{N_1}(1) - Y_{N_1}(0)$
$N_1 + 1$	Regimen 0	$Y_{N_1+1}(1)$	$\mathbf{Y_{N_1+1}(0)}$	$\tau_{N_1+1} = Y_{N_1+1}(1) - Y_{N_1+1}(0)$
⋮	⋮	⋮	⋮	⋮
$N_1 + N_0$	Regimen 0	$Y_{N_1+N_0}(1)$	$\mathbf{Y_{N_1+N_0}(0)}$	$\tau_{N_1+N_0} = Y_{N_1+N_0}(1) - Y_{N_1+N_0}(0)$

clinical trial, is completely characterized by $\mathcal{T} = \{\tau_{(i)},\ i = 1, \ldots, N_1 + N_0\}$. Unfortunately, elements in $\mathcal{T}$ are almost never completely observed, so statistical inference is needed.

Outline of the article

Our primary goal in this article is to provide a brief overview of two classical causal null hypotheses, Fisher's sharp null hypothesis and Neyman's weak null hypothesis, and introduce a novel class of null hypotheses regarding the quantiles of individual treatment effects. We then review some recently proposed methods that conduct exact, randomization-based inference for ITE quantiles. These methods are suitable for a range of practical scenarios, including completely randomized trials, block randomized trials, and a matched-pair design that compares two non-randomized vaccine arms or a new vaccine arm against historical controls. We argue that ITE quantiles, and more generally the distribution function of ITEs, offer a perspective that could complement usual estimands (e.g., sample average treatment effect) in early phase vaccine trials that aim to assess vaccine-induced immunogenicity. Finally, we present a comprehensive case study of the immunogenicity data derived from HIV Vaccine Trials Network (HVTN) Study 086 and explore how different methods may facilitate decision-making and improve the evaluation of vaccine regimens.

Framework, notation, and different null hypotheses

Notation

We consider a two-arm randomized trial on N study participants under Neyman-Rubin's potential outcomes framework [6, 7]. Let $Y_i(1)$ and $Y_i(0)$ denote participant i's potential outcomes under Regimen 1 and Regimen 0 and $\tau_i = Y_i(1) - Y_i(0)$ participant i's ITE. We collect the set of N potential outcomes under Regimen 1 in $\mathbf{Y}(1) = (Y_1(1), Y_i(1), \ldots, Y_N(1))^\intercal$, those under Regimen 0 in $\mathbf{Y}(0) = (Y_1(0), Y_i(0), \ldots, Y_N(0))^\intercal$, and the set of ITEs in $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_N)^\intercal$. Let Z_i denote the treatment assignment to participant i such that $Z_i = 1$ if participant i is assigned Regimen 1 and 0 otherwise. The set of treatment assignments is collected in $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)^\intercal$. For each study participant i, the observed outcome Y_i satisfies $Y_i = Z_i \cdot Y_i(1) + (1 - Z_i) \cdot Y_i(0)$. We collect N observed outcomes in the vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^\intercal$. In randomization inference, potential outcomes of study participants are viewed as fixed quantities and researchers rely *solely* on the treatment assignment mechanism to draw valid causal conclusions. In other words, randomization forms what Sir Ronald Fisher referred to as the "reasoned basis" for causal inference [8]. In a completely randomized experiment (CRE), N_1 study participants are randomly assigned to Regimen 1 and the other $N_0 = N - N_1$ to Regimen 0.

Fisher's sharp null hypothesis and Neyman's weak null hypothesis

Fisher's sharp null hypothesis [8] considers testing H_{δ}: $\boldsymbol{\tau}=\boldsymbol{\delta}$, where $\boldsymbol{\delta} = \left(\delta_1, \delta_2, \ldots, \delta_N\right)^{\intercal} \in \mathbb{R}^N$ is a prespecified vector of constants. The Fisher randomization test (FRT) scheme is often used to test H_{δ}. Following Rosenbaum [9], we consider test statistics of the form $t\left(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\delta}(0)\right)$, where $\boldsymbol{Z}$ is the vector of observed treatment assignments and $\boldsymbol{Y}_{\boldsymbol{Z},\delta}(0)$ is the vector of imputed potential outcomes under Regimen 0. Given the observed data $\boldsymbol{Y}$ and under H_{δ}, we have $\boldsymbol{Y}_{\boldsymbol{Z},\delta}(0)=\boldsymbol{Y}-\boldsymbol{Z}\circ\boldsymbol{\delta}$, where $\circ$ represents element-wise multiplication. In a CRE design, the null distribution of the test statistic has the following tail probability:

$$G_{\boldsymbol{Z},\delta}(c) \equiv \Pr\{t(\boldsymbol{A}, \boldsymbol{Y}_{\boldsymbol{Z},\delta}(0)) \geq c\} = \binom{N}{N_1}^{-1} \cdot \sum_{\boldsymbol{a}\in\{0,1\}^N : \sum_{i=1}^{N} a_i = N_1} \mathbb{1}\left\{t(\boldsymbol{a}, \boldsymbol{Y}_{\boldsymbol{Z},\delta}(0)) \geq c\right\}, \tag{1}$$

where $\boldsymbol{A}$ denotes a random treatment assignment vector under the CRE design, $\boldsymbol{a}$ denotes a realization, and the corresponding randomization-based, exact p-value is obtained by evaluating the tail probability in (1) at the observed value of the test statistic:

$$p_{\boldsymbol{Z},\delta} \equiv G_{\boldsymbol{Z},\delta}\left\{t\left(\boldsymbol{Z}, \boldsymbol{Y}_{\boldsymbol{Z},\delta}(0)\right)\right\}. \tag{2}$$

In some circumstances, treatment effect heterogeneity is likely to exist although its details (e.g., how the treatment effect varies across subgroups) may be unknown. In these cases, testing the following weak null hypothesis of no sample average treatment effect emerges as an alternative [6]:

$$H_{0,\text{weak}}: \overline{\tau} = \frac{1}{N}\sum_{i=1}^{N} \tau_i = 0.$$

Unlike the FRT that is *exact*, randomization-based tests for SATE require large-sample approximation, although some of them also enjoy finite-sample validity for a certain sharp null hypothesis [10–12].

As discussed in Section 1.1, the SATE is one important, albeit not comprehensive, assessment of a regimen's treatment effect profile. Moreover, when the outcomes have heavy tails, the SATE could be sensitive to the outliers and the finite population asymptotic approximation tends to work poorly in these cases [4]. These aspects are particularly relevant in pre-clinical studies (e.g., nonhuman primates studies) and early phase clinical trials, where the sample size could be as small as 10–20 per arm and treatment effect heterogeneity is often expected.

Beyond the SATE: quantiles and proportions

An exclusive focus on the SATE could arrive at unsatisfactory conclusions in the presence of large treatment effect heterogeneity. For example, a treatment that is harmful to a majority of participants could still have a positive treatment effect on average, due to some large ITEs on a small proportion of the cohort.

Some researchers propose to study the difference between quantiles of the marginal distributions of potential outcomes:

$$q_{\Delta} \equiv q_{1,k} - q_{0,k},$$

where $q_{j,k} \equiv \inf_q \Pr\{Y(j) \leq q\} \geq k$, $j=0,1$. These works often adopt a superpopulation setting, from which study participants are drawn as an independent and identically distributed sample. Firpo [13] proposes a semiparametric estimation procedure targeting q_{Δ} under the no unmeasured confounding assumption. Frölich and Melly [14] extend this work to estimating q_{Δ} among compliers using a binary instrumental variable in the presence of unmeasured confounding. More recently, Powell [15] generalizes the so-called generalized quantile regression (GQR) method to estimating q_{Δ} with one or more, possibly continuous, treatment variables.

A comprehensive summary of ITE profile is its distribution function. In this paper, we propose to study the quantiles of ITEs, which involve the joint distribution of $Y(1)$ and $Y(0)$ and differ from the difference in two quantile functions. In fact, q_{Δ} only coincides with the quantile of ITEs when the rank of the potential outcome for a given individual is the same regardless of his or her treatment status [13]. Under a superpopulation setting,

some authors, see, e.g., Fan and Park [16, 17] and references therein, studied sharp bounds on the quantile of ITEs and establish their asymptotic properties. More recently, Huang et al. [18] proposes a pointwise consistent confidence set for the proportion of units benefiting from the treatment for a binary or ordinal outcome. In the rest of this article, including methods reviewed in Section 3 and studied in Section 4, all focus on a *finite sample* setting where the potential outcomes of all study participants are fixed, as opposed to the superpopulation inference which could work poorly in early phase clinical trials with small sample sizes.

Below, we formally introduce our target causal estimand and associated null hypotheses. Let $F(c) \equiv N^{-1}\sum_{i=1}^{N}\mathbb{1}(\tau_i \le c)$, for $c \in \mathbb{R}$, denote the distribution function of ITEs and $F^{-1}(\beta)\equiv\inf\{c\colon F(c)\ge\beta\}$, for $\beta \in (0, 1]$, denote the corresponding quantile function. We focus on the N study participants so that the quantile function can take at most N values, i.e., the sorted ITEs $\tau_{(1)}\le\tau_{(2)}\le \ldots \le\tau_{(N)}$. More precisely, we have $F^{-1}(\beta)=\tau_{(k)}$, with $k=\lceil N\beta\rceil$ denoting the ceiling of $N\beta$, for $\beta \in (0, 1]$. In addition, the distribution function can be equivalently written as $F(c)=1-N(c)/N$, for $c \in \mathbb{R}$, where $N(c) \equiv \sum_{i=1}^{N}\mathbb{1}(\tau_i > c)$ denotes the number of participants with ITEs exceeding a threshold c. We consider the following null hypothesis for any $0\le k\le N$ and $c \in \mathbb{R}$:

$$H_{k,c}\colon \tau_{(k)} \le c \iff N(c) \le N - k. \tag{3}$$

For descriptive simplicity, we define $\tau_{(0)}=-\infty$. Here we focus on one-sided testing with alternatives favoring large treatment effects; by inverting the tests, we may then obtain lower confidence limits for ITE quantiles $F^{-1}(\beta)$'s (or equivalently $\tau_{(k)}$'s) and proportions of participants with ITEs exceeding any thresholds $1-F(c)$'s (or equivalently $N(c)$'s). To obtain one-sided tests with alternatives favoring small treatment effects, one may use the same procedure but with observed outcomes' signs flipped or the treatment/control status switched. Two-sided tests can be constructed by combining two one-sided tests using, say, the Bonferroni correction.

A review of randomization inference for ITE profiles

Completely randomized experiments

In a recent article, Caughey et al. [4] extend the FRT to testing the null hypothesis $H_{k,c}$ in (3). Because $H_{k,c}$ is a composite null hypothesis and permits infinitely many imputation schemes, FRT is not directly applicable. Nevertheless, a valid p-value for testing $H_{k,c}$ can be obtained by maximizing the randomization p-value $p_{\boldsymbol{Z},\boldsymbol{\delta}}$ in (2) over $\boldsymbol{\delta} \in \mathcal{H}_{k,c}$, where $\mathcal{H}_{k,c}$ denotes the set of vectors whose elements of rank k are bounded by c, i.e., $\mathcal{H}_{k,c} \equiv \left\{\boldsymbol{\delta} \in \mathbb{R}^N\colon \delta_{(k)} \le c\right\} \subset \mathbb{R}^N$. However, optimizing $\sup_{\boldsymbol{\delta}\in\mathcal{H}_{k,c}} p_{\boldsymbol{Z},\boldsymbol{\delta}}$ is computationally challenging and may be NP-hard. To address this challenge, Caughey et al. [4] propose to use the class of rank score statistics of the following form:

$$t(\boldsymbol{z},\boldsymbol{y}) = \sum_{i=1}^{N} z_i\phi\{\mathrm{r}_i(\boldsymbol{y})\}, \tag{4}$$

where $\phi\{\cdot\}$ is a monotone increasing function, and $\mathrm{r}_i(\boldsymbol{y})$ denotes the rank of the ith coordinate of $\boldsymbol{y}$ using index ordering to break ties, assuming that the ordering has been randomly permuted before the analysis.

Under a CRE design, the rank score statistic $t(\cdot, \cdot)$ defined in (4) is distribution free, in the sense that for any $\boldsymbol{y},\boldsymbol{y}' \in \mathbb{R}^N$, $t(\boldsymbol{Z}, \boldsymbol{y})$ and $t(\boldsymbol{Z}, \boldsymbol{y}')$ follow the same distribution. Because of this distribution free property, the imputed randomization distribution in (1) reduces to a distribution that does not depend on the observed treatment assignment $\boldsymbol{Z}$ or the hypothesized treatment effect $\boldsymbol{\delta}$, i.e.,

$$G_{\boldsymbol{Z},\boldsymbol{\delta}}(c) \equiv \Pr\left\{t\left(\boldsymbol{A}, \boldsymbol{Y}_{\boldsymbol{Z},\boldsymbol{\delta}}(0)\right) \ge c\right\} = \Pr\{t(\boldsymbol{A},\boldsymbol{y}) \ge c\} \equiv G(c), \tag{5}$$

where $\boldsymbol{y}$ denotes an arbitrary vector in $\mathbb{R}^N$. Consequently, the valid p-value $\sup_{\boldsymbol{\delta}\in\mathcal{H}_{k,c}} p_{\boldsymbol{Z},\boldsymbol{\delta}}$ for testing $H_{k,c}$ in (3) simplifies to

$$p^{\mathrm{R}}_{k,c} \equiv \sup_{\boldsymbol{\delta}\in\mathcal{H}_{k,c}} p_{\boldsymbol{Z},\boldsymbol{\delta}} = \sup_{\boldsymbol{\delta}\in\mathcal{H}_{k,c}} G\{t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\circ\boldsymbol{\delta})\} = G\left\{\inf_{\boldsymbol{\delta}\in\mathcal{H}_{k,c}} t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\circ\boldsymbol{\delta})\right\}, \tag{6}$$

where the last equality holds because G is monotone decreasing and $t(\boldsymbol{Z}, \boldsymbol{Y} - \boldsymbol{Z}\circ\boldsymbol{\delta})$ achieves its infimum over $\boldsymbol{\delta} \in \mathcal{H}_{k,c}$. Equation (6) suggests that, to obtain a valid p-value for testing $H_{k,c}$ based on a distribution free test statistic, it suffices to minimize the value of the test statistic $t\big(\boldsymbol{Z}, \boldsymbol{Y}_{Z,\boldsymbol{\delta}}(0)\big)$ over $\boldsymbol{\delta} \in \mathcal{H}_{k,c}$. When using the rank score statistic in (4), the infimum is achieved when the ranks of $Y_i(0)$'s of treated participants are minimized, or equivalently, treated participants' ITEs are maximized subject to $H_{k,c}$. Because Z_i=0 for those in the control group, their ITEs do not directly contribute to the value of the test statistic. Caughey et al. [4] show that the worst-case p-value corresponds to assigning arbitrarily large ITEs to the $N - k$ treated participants with the largest observed outcomes and c to the remaining participants. Moreover, the inference is simultaneously valid in the sense that there is no need to conduct multiple analysis correction when jointly inferring many quantiles or the entire distribution function of ITEs.

Caughey et al.'s [4] approach can be conservative, since the worst-case consideration corresponds to assigning largest ITEs to the treated participants; this is unlikely by randomization because ITE can be understood as a pretreatment variable in a broad sense, and randomization tends to balance the distribution of pretreatment variables. More recently, Chen and Li [19] propose two enhanced methods that tackle this limitation and can achieve improved statistical power.

Their first method better leverages the information contained in the control participants by (A1) conducting level-$(1 - \alpha)$ simultaneous inference for ITEs among treated participants, (A2) flipping treated and control participants and repeating the level-$(1 - \alpha)$ simultaneous inference for control participants, and (A3) combining the intervals for all participants by ordering the one-sided intervals according to their lower confidence limits. The resulting ordered intervals can be shown to be simultaneously valid, level-$(1 - 2\alpha)$ confidence intervals for all N ITEs. In this procedure, the choice of the test statistic could be different in Step (A1) and Step (A2). We will explore the choice of test statistics in some practical settings in the simulation study. In their second method, instead of presuming that the largest $N - k$ ITEs are all among the treated participants, they view the number of participants with the largest $N - k$ effects in the treatment arm as a nuisance parameter and use Berger and Boos's [20] approach to control for the randomness of this nuisance. Their second method can be summarized as follows: (B1) apply the Berger and Boos's [20] correction to derive simultaneous confidence intervals for all participants; (B2) flip the role of treated and control participants and repeat Step (B1); and (B3) combine the intervals derived from Step (B1) and Step (B2) using the Bonferroni method. Chen and Li [19] show via simulations that both approaches deliver more powerful statistical inference compared to the original method in Caughey et al. [4] while remaining exactly valid in finite sample.

Stratified randomized experiments

Su and Li [21] extend the approach in Caughey et al. [4] to stratified completely randomized experiments (SCRE). They consider testing the same null hypothesis $H_{k,c}$ in a study design with S strata. Each stratum consists of n_s study participants such that $N = \sum_{s=1}^{S} n_s$. Su and Li [21] consider using the following stratified rank sum statistic:

$$t_{\text{str}}(\boldsymbol{z}, \boldsymbol{y}) = \sum_{s=1}^{S} t_s\big(\boldsymbol{z}_{\boldsymbol{s}}, \boldsymbol{y}_{\boldsymbol{s}}\big) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} z_{si}\phi_s\big\{\mathrm{r}_i\big(\boldsymbol{y}_{\boldsymbol{s}}\big)\big\}, \tag{7}$$

where $\phi_s\{\cdot\}$ again denotes some monotone increasing rank transformation for stratum S=s, $\boldsymbol{z}_s$ and $\boldsymbol{y}_s$ denote the subvectors of $\boldsymbol{z}$ and $\boldsymbol{y}$ corresponding to stratum S=s, and $\mathrm{r}_i(\boldsymbol{y}_{\boldsymbol{s}})$ denotes the rank of the ith coordinate of $\boldsymbol{y}_{\boldsymbol{s}}$ among all coordinates of $\boldsymbol{y}_{\boldsymbol{s}}$. Again, the stratified rank sum statistic enjoys the distribution free property in a SCRE, so that a valid p-value for testing $H_{k,c}$ has an equivalent form as in (6). Specifically, the p-value $p^{\mathrm{R}}_{k,c}$ defined as in (6), but with $t(\cdot, \cdot)$ replaced by the stratified rank sum statistic $t_{\text{str}}(\cdot, \cdot)$ and $G(\cdot)$ being the tail probability of $t_{\text{str}}(\boldsymbol{Z}, \boldsymbol{y})$ for any $\boldsymbol{y} \in \mathbb{R}^N$, is a valid p-value for testing $H_{k,c}$. Su and Li [21] then demonstrate that minimizing the test statistic over all possible ITEs compatible with the null hypothesis $H_{k,c}$ can be transformed into a multiple-choice knapsack problem, which can be solved exactly using dynamic programming, or in a slightly conservative manner using a greedy algorithm. Implementation of the methods can be found in the R package QIoT [21]. The first enhanced method in Section 3.1 can also be extended to the SCRE [19].

Integrated analysis of non-randomized arms

In many circumstances, researchers may be interested in conducting a pooled analysis of data derived from multiple clinical trials. This can happen in two circumstances. First, researchers may be interested in comparing immunogenicity of two vaccine regimens, one from a current trial and the other from a historical trial. Second, researchers may be interested in comparing a vaccine regimen to historical controls. Because of a lack of randomization, a naïve comparison of outcomes from non-randomized arms assuming randomization could lead to a bias in estimating the treatment effects. In these studies, it is essential to adjust for baseline covariates that could predict immunogenicity; for instance, Huang et al. [2] report that host baseline characteristics could predict a high-level binding antibody response with a cross-validated area under the receiver operating characteristics curve (AUROC) equal to 0.72 (95 % CI: [0.68 0.76]). Many methods can be used to perform covariance adjustment; in early phase trials with small samples, one reasonable strategy is to conduct a matched cohort study [9, 22]. One popular downstream data analysis strategy in a matched cohort study is to conduct randomization inference. Because of the matched-pair or matched-set structure, one needs to conduct stratified randomization inference discussed in Section 3.2. We will illustrate the proposed workflow, from statistical matching to stratified randomization inference for the ITE profile, in a case study in Section 6.

Relaxing the randomization assumption

In a matched analysis of immunogenicity data derived from non-randomized arms, the key assumption is a version of the ignorability assumption, which effectively says that within the strata defined by observed covariates being matched on, selection into a particular trial or study arm is randomized [23, 24]. In these analyses, it is essential to examine the consequences of deviating from the randomization assumption, as Rosenbaum [25] (Chapter 6) put it: "the absence of an obvious reason to think that two groups are different falls well short of a compelling reason to think they are the same." In early phase vaccine trials, it is conceivable that healthier study participants with potentially stronger immune responses may be preferentially enrolled a certain trial or study arm, for instance, because of the difference in the vaccine dosing schedules (bolus vs. fractional dosing), and adjusting for observed covariates may not be sufficient in removing the selection bias.

Su and Li [21] study a relaxed randomization inference scheme, where in lieu of assuming randomization within each stratum, the treatment assignment probability is allowed to deviate from randomization under a model that controls the maximum level of deviation [9, 26]:

$$\frac{1}{\Gamma} \leq \frac{\pi_{sj}/(1-\pi_{sj})}{\pi_{sj'}/(1-\pi_{sj'})} \leq \Gamma, \quad 1 \leq j, j' \leq n_s,\ 1 \leq s \leq S, \tag{8}$$

where π_{sj} and $\pi_{sj'}$ denote the treatment assignment probabilities of participant j and j' in stratum $S{=}s$, respectively, and the sensitivity parameter $\Gamma{\geq}1$ specifies the maximum odds ratio across all strata. Inferring the treatment effect under a biased randomization scheme is referred to as a sensitivity analysis in the analysis of non-randomized or observational data, and the goal is to investigate the maximum degree of deviation from the randomization assumption when a certain null hypothesis can no longer be rejected. Such sensitivity analysis methods have been developed for Fisher's sharp null hypothesis [9, 26] and Neyman's weak null hypothesis of no sample average treatment effect [27]; Su and Li [21] generalize the method to testing the null hypothesis $H_{k,c}$ under a matched design.

Placebo-controlled trials with highly specific endpoints

In a placebo-controlled vaccine trial, participants' potential outcomes under the placebo are often known *a priori*, and these controls are nonetheless included in the study primarily for blinding purposes (e.g., preventing treatment arm information from being revealed to lab technicians). For instance, in many early phase HIV vaccine trials, the endpoints of interest are vaccine-induced, antigen-specific immune responses, and healthy and

naïve study participants receiving the placebo are not expected to exhibit any of these highly specific immune responses to the antigens. Hence, we often have auxiliary information $Y_i(0) = \mathsf{LOD}$ for all i=1, …, N, where LOD represents an assay-specific limit of detection. In this stylistic case, we immediately know the ITEs of N_1 treated participants, based on which we may infer the entire ITE distribution. In this section, we formally derive a level-$(1-\alpha)$ confidence interval for a single ITE quantile $\tau_{(k)}$, prove the equivalence between our result and earlier results by Sedransk and Meyer [28] in the context of simple random sampling, and derive simultaneously valid confidence intervals for multiple ITE quantiles or the entire ITE distribution. Methods in this section always leverage the auxiliary information about $Y(0)$ and therefore are often more powerful than those developed for generic endpoints reviewed in Section 3. There are two ways to see why leveraging such auxiliary information makes inference more powerful. First, all $Y_i(0)$'s in the science Table 1 are known *a priori* and the only potential outcomes not known are $Y_i(1)$'s of the control participants. A second and related perspective is to consider the worst-case p-value in expression (6). Auxiliary information on $Y_i(0)$ places restriction on the set $\mathcal{H}_{k,c}$ and hence the supremum is taken over a smaller space.

Without loss of generality, we assume $\mathsf{LOD} = 0$ so $Y_i(0)$=0 for all i's. If $\mathsf{LOD} \neq 0$, then we can always shift $Y(0)$ and $Y(1)$ by the magnitude of LOD and proceed to make inference on the transformed outcomes. ITEs will remain unchanged. Consider the null hypothesis $H_{k,c}$ in (3) for any $1 \le k \le N$ and $c \in \mathbb{R}$. Under the CRE design, treated participants are a simple random sample of size N_1 from a total of N participants. Therefore, $n(c) \equiv \sum_{i=1}^{N} Z_i \mathbb{1}(Y_i > c) = \sum_{i=1}^{N} Z_i \mathbb{1}(\tau_i > c)$ follows a Hypergeometric distribution with parameters $(N, N(c), N_1)$. The Hypergeometric distribution with parameters (N, n, N_1) becomes stochastically larger as n increases; that is, if X follows a Hypergeometric distribution with parameters (N, n, N_1), Y follows a Hypergeometric distribution with parameters $\left(N, n', N_1\right)$ such that $n' \ge n$, then $\mathbb{P}(X \le t) \ge \mathbb{P}(Y \le t)$ for every $t \in \mathbb{R}$. Together, these facts imply a finite-sample valid p-value for testing $H_{k,c}$, as summarized in the following proposition.

Proposition 1. *Consider a CRE design and assume $Y_i(0)$=0 for i=1, …, N. For any $1 \le k \le N$ and $c \in \mathbb{R}$, $p^H_{k,c} \equiv G_H(n(c); N, N-k, N_1)$ is a valid p-value for testing the null hypothesis $H_{k,c}$ in (3), where $G_H(x; N, n, N_1) \equiv Pr(X \ge x)$ denotes the tail probability of a Hypergeometric random variable X with parameters (N, n, N_1). Specifically, under $H_{k,c}$, $\Pr\left(p^H_{k,c} \le \alpha\right) \le \alpha$ for any $\alpha \in (0, 1)$.*

Proof All proofs in the article can be found in Supplementary Material A.

From Proposition 1, we can then conduct Lehmann-style test inversion to construct confidence intervals for $\tau_{(k)}$ and $n(c)$, for any $1 \le k \le N$ and $c \in \mathbb{R}$. Moreover, due to the monotonicity of the p-value $p^{\mathrm{H}}_{k,c}$ in k and c, the resulting confidence sets are intervals and have simpler forms that facilitate their computation. Proposition 2 summarizes these results. For descriptive convenience, we let $y_{(1)} \le y_{(2)} \le \ldots \le y_{(N_1)}$ denote the sorted observed outcomes for treated participants and further define $y_{(0)} = -\infty$.

Proposition 2. *Under a CRE design where $Y_i(0)$=0 for i=1, …, N, we have the following:*

(a) *$p^H_{k,c}$ is monotone increasing in c and decreasing in k.*

(b) *For any $1 \le k \le N$ and $\alpha \in (0, 1)$, $\left\{c: p^H_{k,c} > \alpha, c \in \mathbb{R}\right\}$ is a $1-\alpha$ confidence interval for $\tau_{(k)}$, and it can be equivalently written as $[y_{(k(\alpha))}, \infty)$ with $k(\alpha) \equiv N_1 - Q_H(1-\alpha; N, N-k, N_1)$, where $Q_H(\theta; N, n, N_1)$ denotes the θ-th quantile function of a Hypergeometric random variable X with parameters (N, n, N_1).*

(c) *For any $c \in \mathbb{R}$ and $\alpha \in (0, 1)$, $\left\{N-k: p^H_{k,c} > \alpha, 0 \le k \le N\right\}$ is a $1-\alpha$ confidence set for N(c), and it can be equivalently written as $\{n_{c,\alpha}, n_{c,\alpha}+1, \ldots, N\}$ with $n_{c,\alpha} = N - \max\{k: G_H(n(c); N, N-k, N_1) > \alpha, 0 \le k \le N\}$.*

Confidence interval results for $\tau_{(k)}$ can also be derived by applying results in Sedransk and Meyer [28] who study inference for quantiles of a finite population based on simple random sampling. In Supplementary Material A.3, we prove the equivalence between the confidence intervals in Proposition 2 and those derived by directly applying results in Sedransk and Meyer [28]. In Supplementary Material A.3, we also prove that the confidence intervals for $N(c)$ in Proposition 2 are equivalent to that in Wang [29], who studies inference for

Hypergeometric distribution parameters. Hence, by the optimality results in Wang [29], confidence intervals for $N(c)$ in Proposition 2(c) are optimal one-sided confidence intervals.

Evaluating the ITE profile of a treatment regimen goes beyond constructing confidence intervals for each individual ITE. Proposition 3 derives the simultaneous confidence intervals for multiple ITE quantiles, which provides a valid summary of treatment effect profile and is of primary interest in practice.

Proposition 3. *Consider a CRE design and assume $Y_i(0)=0$ for $i=1,\ldots,N$. Suppose that we are interested in multiple quantiles of ITEs: $\tau_{(k_1)}, \tau_{(k_2)}, \ldots, \tau_{(k_J)}$, where $1\leq k_1\leq \ldots \leq k_J\leq N$. Under a CRE design, the simultaneous coverage probability for quantiles of ITEs is*

$$\Pr\left(\bigcap_{j=1}^{J}\{\tau_{(k_j)} \geq y_{(k_j(\alpha))}\}\right) \geq 1-\Pr\left(\bigcup_{j=1}^{J}\left\{\sum_{i=1}^{N} Z_i \mathbb{1}(i>k_j) > Q_H(1-\alpha; N, N-k_j, N_1)\right\}\right), \tag{9}$$

where the equality holds when all ITEs τ_i's are distinct. Recall that $y_{(k_j(\alpha))}$ is the observed outcome of rank $k_j(\alpha)\equiv N_1 - Q_H(1-\alpha; N, N-k_j, N_1)$ in the treatment group.

Proposition (9) is useful because the lower bound depends only on observed data quantities and thus can be efficiently approximately using Monte Carlo methods. Moreover, the lower bound of the simultaneous coverage probability is sharp in the sense that it is achieved when all τ_i's are distinct across all N participants.

Technically, the assay limit of detection specifies the lowest level of immune response to be detected and merely places an upper bound on $Y_i(0)$ rather than specifies its precise value. Proposition 4 extends previous results and suggests that p-values and confidence statements derived from Propositions 1–3 remain valid when $Y_i(0)\leq 0$ rather than $Y_i(0)=0$.

Proposition 4. *Propositions 1–3 hold when all individuals have a non-positive and possibly different potential outcome under control, that is, $Y_i(0)\leq 0$ for all $i=1,\ldots,N$.*

Simulation

Characterizing ITE profiles with highly specific endpoints

One primary objective in a typical experimental vaccine trial is to *characterize* the immunogenicity profile of each candidate regimen. The primary goal of our first simulation study is to compare the power of pointwise and simultaneous inference of ITE quantiles when $Y_i(0) = \mathsf{LOD}$ using the methodology introduced in Section 4. Our simulation studies can be compactly summarized by a $2\times 2\times 3$ factorial design. Factors 1 and 2 specify the clinical trial design and are meant to mimic the size of a typical early phase or experimental medicine trial:

Factor 1: sample size, N: 40 and 100;

Factor 2: proportion assigned to treatment, $p\equiv N_1/N=0.5$ (balanced design).

We consider a simple pattern for the ITE profile: $\alpha\times N$ units have a null ITE ($\tau_i=0$) and the other $(1-\alpha)\times N$ positive responders have a positive ITE ($\tau_i>0$). The distribution of τ_i among $(1-\alpha)\times N$ positive responders follows a truncated normal distribution with parameters μ, $\sigma=1.5$, and truncation at 0.1 and 4. This simple ITE pattern is meant to mimic a practical setting in HIV vaccine development where the experimental vaccine only induces an antigen-specific immune response for a subset of study participants. Factors 3 and 4 specify α and μ.

Factor 3: proportion of units exhibiting a null ITE, α: 0.2 and 0.5;

Factor 4: mean parameter of truncated normal among positive responders, μ: 1.5, 2.0 and 2.5.

We consider a constant potential outcome under control, $Y_i(0)$=2, corresponding to LOD in our case study. Each unit has its potential outcome $Y_i(1)=Y_i(0)+\tau_i$, and the observed outcome Y_i satisfies $Y_i=Z_i \cdot Y_i(1)+(1-Z_i) \cdot Y_i(0)$.

For each of the 12 data generating processes, we performed randomization inference and constructed (i) individual confidence intervals; and (ii) simultaneous confidence region for selected quantiles (six quantiles including $\lceil 0.5 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$ and $\lceil 0.95 \times N \rceil$, 10 quantiles including $k=\lceil 0.5 \times N \rceil$, $\lceil 0.6 \times N \rceil$, $\lceil 0.65 \times N \rceil$, $\lceil 0.7 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$, $\lceil 0.95 \times N \rceil$ and N, and 15 quantiles including $\lceil 0.3 \times N \rceil$, $\lceil 0.35 \times N \rceil$, $\lceil 0.4 \times N \rceil$, $\lceil 0.45 \times N \rceil$, $\lceil 0.5 \times N \rceil$, $\lceil 0.55 \times N \rceil$, $\lceil 0.6 \times N \rceil$, $\lceil 0.65 \times N \rceil$, $\lceil 0.7 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$, $\lceil 0.95 \times N \rceil$ and N).

For each data generating process, we repeated simulation 1,000 times. We measured the success of each method in recovering the ITE profile according to the following two criteria. First, for a selected quantile k and the target estimand $\tau_{(k)}$, we recorded $\hat{L}_{(k)}$, the 95 % one-sided lower confidence limits, across 1,000 simulations and plot its cumulative distribution function. Second, we calculated $\mathsf{SS} = |\mathcal{K}|^{-1}\sum_{k\in\mathcal{K}}(\hat{L}_{(k)}-\tau_{(k)})^2$, where $\mathcal{K}$ is a collection of quantiles of interest, and $\hat{L}_{(k)}$ was derived using the pointwise or simultaneous inference methods. For non-informative lower confidence limits, we set $\hat{L}_{(k)}$ to -10. We reported SS averaged over 1,000 simulations. A smaller SS value is more preferable.

Figures S1 and S2 in the Supplementary Material B contrast the distributions of $\hat{L}_{(k)}$ across six quantiles derived from pointwise (left panel) and simultaneous inference (right panel). Table 2 summarizes SS for each data generating process. In general, we conclude that the simultaneous inference yielded lower confidence limits only slightly inferior compared to those derived from pointwise inference, especially when the sample size is small and number of quantiles is relatively small.

Compare different methods with generic endpoints

The goal of our second simulation study is to assess and compare the power of several competing methodologies reviewed in Section 3 when comparing two vaccine regimens with no restrictions on the endpoints. To provide more relevant guidance for practice, we will use the immune response data generated from HVTN 086 as a basis for the data generating process. Specifically, we considered the following two data generating processes:

Table 2: Average SS over 1,000 simulations with $Y_i(0)$=2 for various sample size n, proportion of units with a null treatment effect α, and mean of truncated normal among positive responders μ. **Point**: pointwise inference. **Simul**: simultaneous inference over 6, 10 or 15 ITE quantiles.

n	α	μ	6 quantiles		10 quantiles		15 quantiles	
			Point	Simul	Point	Simul	Point	Simul
40	0.20	1.50	0.37	0.37	0.34	0.38	0.38	0.44
40	0.20	2.00	0.37	0.37	0.34	0.38	0.42	0.49
40	0.20	2.50	0.36	0.36	0.33	0.38	0.51	0.60
40	0.50	1.50	0.59	0.59	0.56	0.63	0.41	0.50
40	0.50	2.00	0.60	0.60	0.67	0.76	0.47	0.60
40	0.50	2.50	0.65	0.65	0.77	0.92	0.56	0.73
100	0.20	1.50	0.13	0.18	0.12	0.17	0.13	0.20
100	0.20	2.00	0.12	0.17	0.11	0.16	0.16	0.25
100	0.20	2.50	0.10	0.14	0.10	0.15	0.17	0.28
100	0.50	1.50	0.20	0.28	0.24	0.33	0.18	0.26
100	0.50	2.00	0.20	0.28	0.31	0.41	0.22	0.32
100	0.50	2.50	0.19	0.27	0.34	0.46	0.26	0.39

- **DGP I:** We sampled BAMA responses of size N_1 against Con 6 gp120/B with replacement from the vaccine regimen T1 and of size N_0 against Con 6 gp120/B with replacement from the vaccine regimen T2.
- **DGP II:** We sampled BAMA responses of size N_1 against gp41 with replacement from the vaccine regimen T1 and data of size N_0 against gp41 with replacement from the vaccine regimen T2.

For each of the $N{=}N_1 + N_0$ sampled data points, a Gaussian noise $\epsilon \sim N(0, 0.15^2)$ was added to their $\log_{10}$-transformed scale. DGP I represents a scenario where two vaccine regimens have similar spread in their immune responses and the treatment effects appear homogeneous. DGP II represents a distinct scenario where immune responses from two vaccine regimens have rather different spread and a heterogeneous ITE profile among vaccine recipients seems more plausible. We considered balanced design with the following three sets of sample sizes: $N_1{=}N_0{=}30$, $N_1{=}N_0{=}50$, and $N_1{=}N_0{=}100$.

For each data generating process, we performed randomization inference for the distribution function of ITEs (and equivalently the proportion of units with ITEs exceeding different thresholds). We constructed (i) individual confidence intervals for selected quantiles; and (ii) simultaneous confidence intervals for selected quantiles using the following methods and choices of test statistics:

- M1: The original method in Caughey et al. [4]. We considered using Stephenson rank sum test with $s{=}2$ (M1-S2) and $s{=}6$ (M1-S6);
- M2: The first enhanced method in Chen and Li [19] which combines inference for treated and control participants. This method involves choosing a test statistic when making inference for the treated and a second test statistic when making inference for the control. We considered the following $2 \times 2{=}4$ combinations of test statistics: Stephenson rank sum test with $s{=}2$ or 6 when inferring ITEs for the treated in Step (A1) and Stephenson rank sum test with $s{=}2$ or 6 when inferring ITEs for the control in Step (A2). These four methods are referred to as M2-S2-S2, M2-S2-S6, M2-S6-S2, and M2-S6-S6. For instance, M2-S2-S6 represents the method that uses Stephenson rank sum test with $s{=}2$ in Step (A1) and with $s{=}6$ in Step (A2).

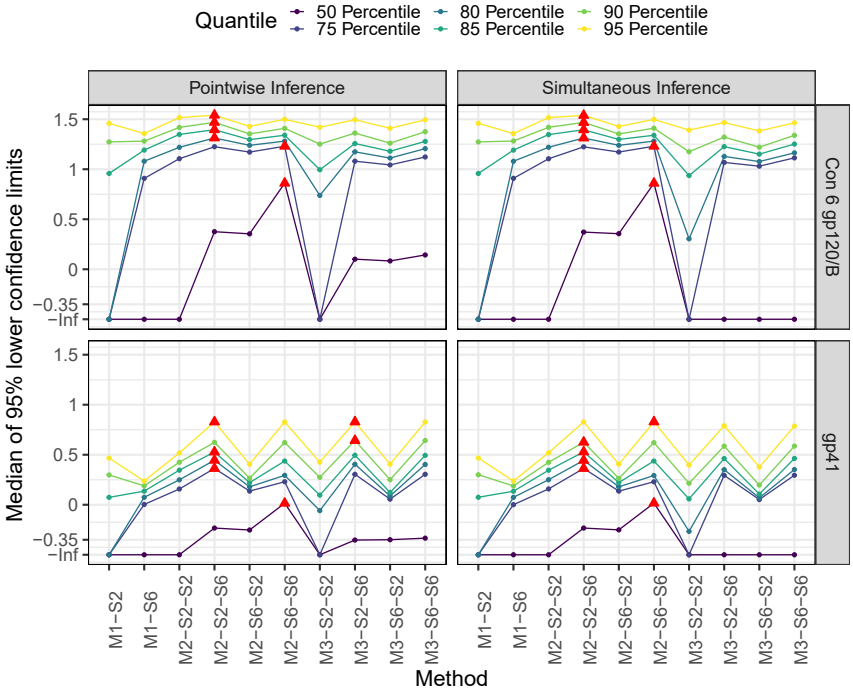


Figure 2: Median of the 95 % lower confidence limits of $\tau_{(k)}$ over 1,000 simulations derived from each method, for **DGP I** (Con 6 gp 120/B) and **DGP II** (gp41) with $N_1{=}N_0{=}30$ and $k{=}\lceil 0.5 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$, and $\lceil 0.95 \times N \rceil$. The red triangle represents the largest median for each quantile.

– M3: The second enhanced method in Chen and Li [19] which considers a probabilistic allocation of the worst-case ITEs. This method also involves making inference twice, once using the raw data and a second time using data with flipped treatment assignments and negated outcomes, before combining the inference. Analogous to M2, we considered the following $2 \times 2 = 4$ combinations of test statistics: Stephenson rank sum test with $s=2$ or 6 when inferring ITEs using raw data in Step (B1) and Stephenson rank sum test with $s=2$ or 6 when inferring ITEs using data with flipped treatment assignments and negated outcomes in Step (B2). These four methods are referred to as M3-S2-S2, M3-S2-S6, M3-S6-S2, and M3-S6-S6.

Figures 2 and 3 plot the median of the 95 % lower confidence limits for selected ITE quantiles when $N_1=N_0=30$ and $N_1=N_0=100$, respectively. Analogous plot when $N_1=N_0=50$ can be found in Supplementary Material B. For each fixed quantile, the red triangle represents the maximum median across different methods. Table 3 further summarizes the average SS of each method under different data generating processes.

We identify several consistent trends from the simulation results. First, each of the three methods exhibits improved power, as reflected by smaller average SS, as the sample size $N=N_1+N_0$ increases. For instance, the average SS drops from 1.09 to 0.78 when N increases from 60 to 200. Second, methods M2 and M3 in general largely outperform M1 in both pointwise and simultaneous inference. In fact, the gain of M2 and M3 over M1 in relatively smaller quantiles like 50 and 75 % quantiles can be quite significant. Third, unlike M1 and M2 whose pointwise and simultaneous inference coincides, M3 suffers from multiple testing correction although the loss in power is negligible when the sample size is only moderately large (e.g., $N_1=N_0=50$). For instance, the average SS for M3-S2-S6 increases from 1.05 to 1.10 when $N_1=N_0=50$.

When we examine different choices of test statistics for M2, methods M2-S2-S6 and M2-S6-S6 have similar performance and both outperform M2-S2-S2 and M2-S6-S2 for both DGP I and DGP II. Both M2-S2-S6 and M2-S6-S6 use the Stephenson rank sum test with a large s, e.g., $s=6$, when the treatment status is flipped (i.e., Step A2 in M2 and Step B2 in M3). This is consistent with the observation that a large s is preferred when treated participants

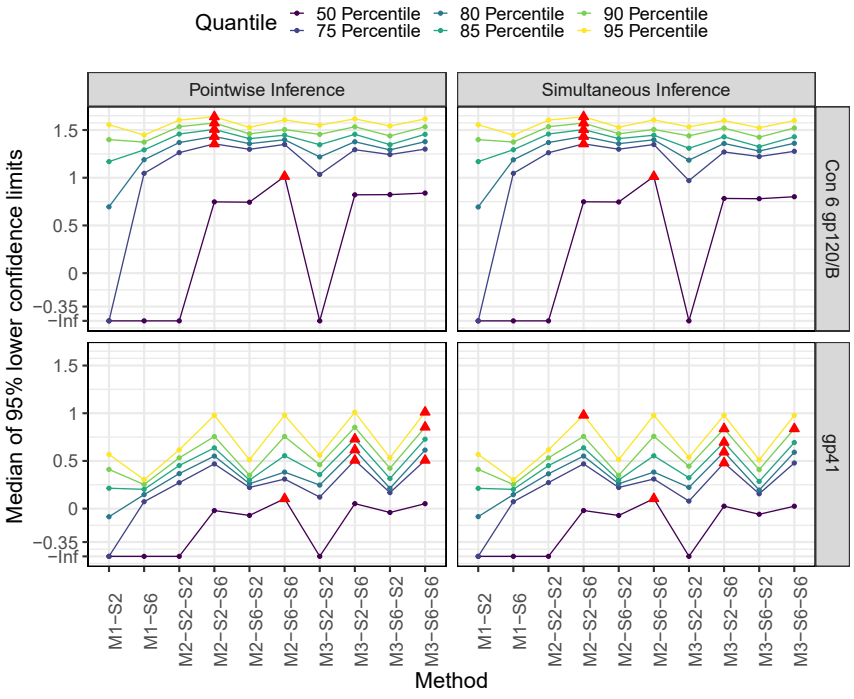


Figure 3: Median of the 95 % lower confidence limits of $\tau_{(k)}$ over 1,000 simulations derived from each method, for **DGP I** and **DGP II** with $N_1=N_0=100$ and $k=\lceil 0.5 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$, and $\lceil 0.95 \times N \rceil$. The red triangle represents the largest median for each quantile.

Table 3: SS averaged over 1,000 simulations derived from each method for different data generating processes.

DGP	N_1	N_0	Inference	M1-S2	M1-S6	M2-S2-S2	M2-S2-S6	M2-S6-S2	M2-S6-S6	M3-S2-S2	M3-S2-S6	M3-S6-S2	M3-S6-S6
DGP I	30	30	Pointwise	73.41	24.27	23.96	1.09	1.26	0.96	48.84	2.23	2.63	1.36
DGP I	30	30	Simultaneous	73.41	24.27	23.96	1.09	1.26	0.96	49.26	24.10	24.25	24.05
DGP I	50	50	Pointwise	49.10	24.19	23.88	0.92	1.08	0.87	24.30	1.05	1.21	1.02
DGP I	50	50	Simultaneous	49.09	24.19	23.88	0.92	1.07	0.87	24.47	1.10	1.26	1.07
DGP I	100	100	Pointwise	48.78	24.10	23.81	0.78	0.92	0.76	24.01	0.82	0.96	0.81
DGP I	100	100	Simultaneous	48.78	24.10	23.81	0.78	0.92	0.76	24.07	0.87	1.00	0.86
DGP II	30	30	Pointwise	64.92	21.67	21.15	1.31	2.01	1.37	43.09	2.30	2.74	1.39
DGP II	30	30	Simultaneous	64.92	21.67	21.16	1.31	2.01	1.37	43.36	20.81	21.59	20.81
DGP II	50	50	Pointwise	43.19	21.62	21.06	1.18	1.90	1.27	21.44	1.14	1.97	1.14
DGP II	50	50	Simultaneous	43.16	21.62	21.06	1.18	1.90	1.27	21.56	1.20	2.03	1.20
DGP II	100	100	Pointwise	43.04	21.57	20.98	1.05	1.79	1.17	21.19	0.92	1.77	0.92
DGP II	100	100	Simultaneous	43.04	21.57	20.98	1.05	1.79	1.17	21.25	0.97	1.82	0.97

(or in this case, control participants before flipping the treatment status) have right-skewed outcomes and/or many large outliers [4]. Similar observation stands for method M3.

Application to early phase HIV-1 vaccine trials

HVTN 086

Heterogeneity in vaccine-induced immune responses has been widely observed among vaccinees. For instance, as shown in Figure 1, vaccine-induced immune responses are heterogeneous among participants receiving four vaccine regimens in the HVTN 086 study, spanning from participants with no response ("nonresponders") to those with exceptional responses ("high-responders"). Recently, Huang et al. [2] conducted a comprehensive meta-analysis exploring the variations in immune responses induced by 26 vaccine regimens. We focus on the immunogenicity data derived from HVTN 086, a multicenter, randomized, placebo-controlled phase-1 clinical trial studying four vaccine regimens: SAAVI MVA-C priming with sequential or concurrent Novartis subtype C gp140/MF59 vaccine boost and SAAVI DNA-C2 priming with SAAVI MVA-C boosting, with or without Novartis subtype C gp140/MF59 vaccine. The study comprised 4 study arms. Each arm planned to enroll 46 HIV-negative, healthy, vaccine-naïve adult participants between 18 and 45 in the Republic of South Africa, among whom 38 were randomly assigned to one candidate vaccine regimen and the other eight to the placebo. Study participants were randomized to one of the four study arms, and within each study arm, further randomized to the active vaccine regimen or placebo. Below, we follow Huang et al. [2] and consider data from study participants who successfully completed all study visits and received all scheduled vaccination. Figure 4 summarizes scenarios determined by the nature of the study (randomized or not randomized), type of endpoints ($Y_i(0) \leq$LOD or not), and the experimental design (CRE or SCRE) and how they are related to methods reviewed or developed in the article.

Characterizing immunogenicity profiles against placebo

Our first goal is to characterize each vaccine regimen's immunogenicity profile of antigen-specific immune responses within each study arm. The primary outcome of interest is the serum IgG response to the antigen Con6 gp120/B measured by a validated binding antibody multiplex assay (BAMA) two weeks post the last vaccination. Because study participants were all naïve to the antigen, it is reasonable to assume that their potential outcomes under placebo were less than or equal to 100, the limit of detection of the BAMA assay.

Figure 5 plots the 95 % one-sided confidence intervals for selected ITE quantiles using methods described in Section 4. The vertical dashed red lines indicate the 95 % lower confidence limits of the sample average treatment effect derived from a randomization-based test based on the t-statistic [6, 30]. Inference for the effect quantiles

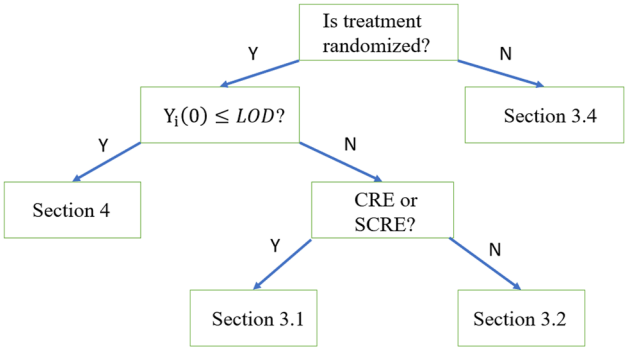
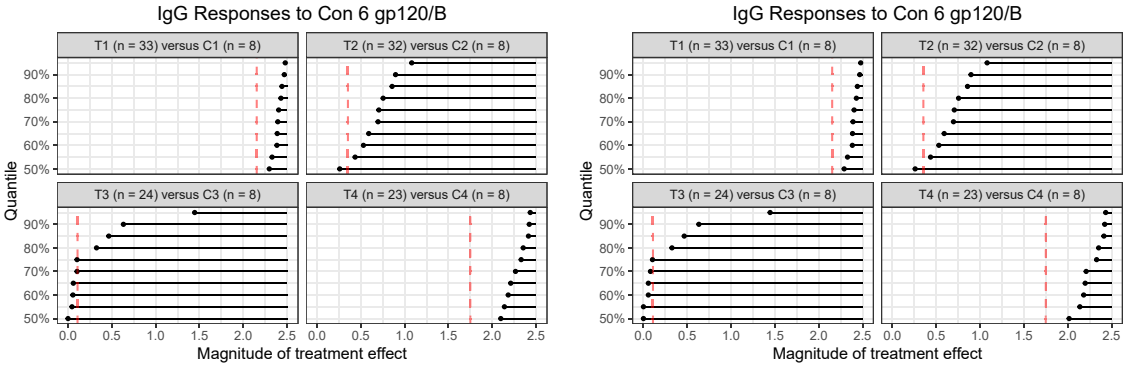


Figure 4: A flow chart relating each study scenario to the methods reviewed or developed in the article.



(a) Inference for a single quantile of treatment effects

(b) Simultaneous inference for multiple quantiles of treatment effects

Figure 5: Vaccine regimens vs. placebo: 95 % (a) pointwise and (b) simultaneous one-sided confidence intervals of $\tau_{(k)}$'s for k=$\lceil 0.5 \times N \rceil$, $\lceil 0.55 \times N \rceil$, $\lceil 0.6 \times N \rceil$, $\lceil 0.65 \times N \rceil$, $\lceil 0.7 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$, and $\lceil 0.95 \times N \rceil$. (a) Inference for a single quantile of treatment effects. (b) Simultaneous inference for multiple quantiles of treatment effects.

largely enriches the data summary based on the SATE; for instance, in addition to stating that the lower confidence limit of SATE of the vaccine regimen T1 (comparing to the placebo) is 2.15 (in the $\log_{10}$-scale), researchers could further report at a specified confidence level that the largest ITE is at least 2.49, the top 25 % ITE is at least 2.40, and the median ITE is at least 2.30, etc. Moreover, unlike the inference for the SATE that requires large-sample approximation, inference for the quantiles is *exact*. Figure 5 also suggests that the simultaneous confidence intervals are very similar to pointwise confidence intervals, suggesting that conducting simultaneous inference does not sacrifice much power.

Inferred ITE profiles also facilitate comparing and ranking four candidate regimens via making inference for $N(c)$, the number of participants with ITEs exceeding a given threshold c. Table 4 summarizes the 95 % lower confidence limits of $N(c)$ for each regimen and selected values of c. According to Table 4, a 95 % confidence interval for $N(2)$ is [31, 41] for T1, suggesting that we are 95 % confident that at least 31/41=75.6 % participants had a treatment effect as large as 2 (in the $\log_{10}$-scale). Similarly, we are 95 % confident that at least 17/31=54.8 % participants had a treatment effect as large as 2 (in the $\log_{10}$-scale) for regimen T4. Based on these results, if researchers are interested in advancing a vaccine regimen that can elicit high immune responses among a large proportion of participants, then T1 is most promising based on data generated from this early phase clinical trial.

Head-to-head comparisons of two vaccine regimens

Having a placebo arm and an *a priori* known $Y(0)$ is a favorable scenario. Below, we consider a direct, head-to-head comparison of three pairs of vaccine regimens in HVTN 086: T1 vs. T2, T1 vs. T3, and T1 vs. T4. A head-to-head comparison of two active vaccine regimens is useful because in most recent early phase vaccine trials, a placebo arm is no longer employed and study participants are often randomized to different active vaccine regimens.

Table 4: Pointwise 95 % lower confidence limits of $N(c)$ for four vaccine regimens, with c=0, 0.5, 1, 1.5 and 2.

Regimen/c	0	0.5	1	1.5	2
T1 (n=33) vs. C1 (n=8)	40	40	40	40	31
T2 (n=32) vs. C2 (n=8)	27	17	3	1	0
T3 (n=24) vs. C3 (n=8)	15	4	3	0	0
T4 (n=23) vs. C4 (n=8)	29	29	24	23	17

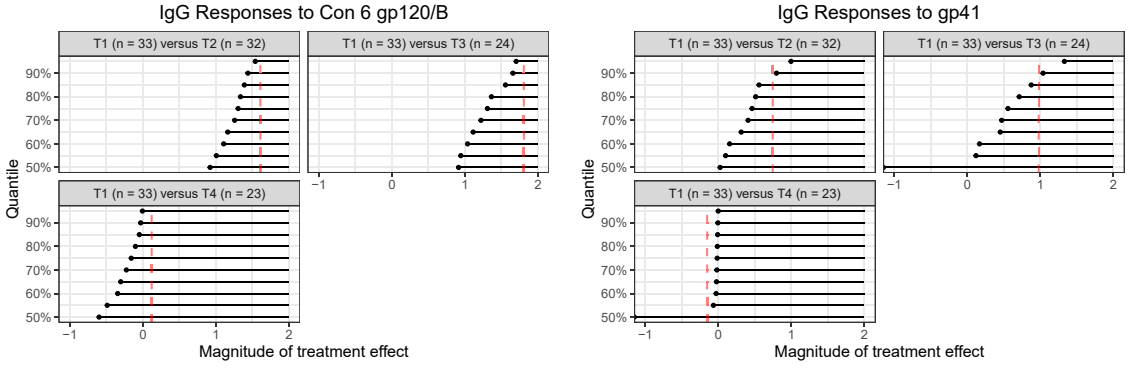
Figure 6 shows the 95 % one-sided simultaneous confidence intervals for selected ITE quantiles of regimen T1 vs. the other three regimens. The vertical dashed red lines show the 95 % lower confidence limits of the SATE derived from a randomization test based on the t-statistic. For both Con 6 gp120/B and gp41, we conducted ITE inference with the method M2. Specifically, we chose the Stephenson rank sum statistic with s=6 in both Step (A1) and Step (A2) for Con 6 gp120/B; as for gp41, we again used the Stephenson rank sum statistic but with s=2 in Step (A1) and s=6 in Step (A2). These choices were guided by the simulation studies in Section 5.

According to Figure 6(b), the 95 % lower confidence limit for the SATE comparing T1 vs. T2 is 0.74 for gp41. Inference for ITE further reveals treatment effect heterogeneity of T1 vs. T2: we are 95 % confident that more than 12.3 % of participants benefited more from T1 compared to T2 by at least 0.74 in the $\log_{10}$-scale (or equivalently a 5.50 times increase in the raw readout) and, at the same time, at least 4.6 % participants benefited by more than 1 unit in the $\log_{10}$-scale (or equivalently 10 times) and 1.5 % benefited by 1.25 (or equivalently 17.78 times). In fact, because the inference for ITE is simultaneously valid, we can immediately reject a constant additive treatment effect of 0.74 for all participants at 95 % confidence level. Analogous inference can be made when comparing T1 to T3.

On the other hand, according to Figure 6(a), the confidence intervals for ITE quantiles comparing T1 vs. T2 for Con 6 gp120/B cover 1.61, the lower confidence limit of the SATE, simultaneously. For Con 6 gp120/B, a constant additive treatment effect of 1.61 for all participants is in fact compatible with the observed data and cannot be rejected at 95 % confidence interval. Compared to gp41, response to Con 6 gp120/B appears to be less heterogeneous when comparing T1 vs. T2 or T3. These conclusions are consistent with the visual display of the data in Figure 1.

A matched study of non-randomized arms

We next considered a head-to-head, across-trial comparison of two regimens – regimen T1 from HVTN 086 (n=33) and regimen T4 from HVTN 205 (n=60) – based on the serum IgG response to antigen gp41 [2]. HVTN 205 was a phase 2a study designed to evaluate DNA and recombinant modified vaccinia Ankara (MVA62B) vaccines [31]. To alleviate the "trial selection bias," we used statistical matching to control for observed baseline characteristics. Specifically, we used an optimal tripartite matching method [32] and constructed 33 matched pairs, each consisting of one study participant receiving vaccine regimen T1 from HVTN 086 and the other receiving T4 from HVTN 205. The matching algorithm closely matched on 11 baseline covariates and minimized the earth mover's distance between the estimated propensity score distributions of two groups. Table 5 summarizes the covariate balance



(a) Simultaneous inference for Con 6 gp120/B using M2-S6-S6 **(b)** Simultaneous inference for gp41 using M2-S2-S6

Figure 6: Head-to-head comparison of vaccine regimens in terms of the binding antibody response to (a) Con 6 gp120/B and (b) gp41: 95 % simultaneous one-sided confidence intervals of $\tau_{(k)}$'s for k=$\lceil 0.5 \times N \rceil$, $\lceil 0.55 \times N \rceil$, $\lceil 0.6 \times N \rceil$, $\lceil 0.65 \times N \rceil$, $\lceil 0.7 \times N \rceil$, $\lceil 0.75 \times N \rceil$, $\lceil 0.8 \times N \rceil$, $\lceil 0.85 \times N \rceil$, $\lceil 0.9 \times N \rceil$, and $\lceil 0.95 \times N \rceil$. (a) Simultaneous inference for Con 6 gp120/B using M2-S6-S6. (b) Simultaneous inference for gp41 using M2-S2-S6.

Table 5: Covariate balance before and after matching. SMD=standardized mean difference.

Covariates	Before matching			After matching	
	HVTN 086 T1 (n=33)	HVTN 205 T4 (n=60)	SMD	HVTN 205 T4 (n=33)	SMD
Age, years	23.6	26.5	−0.37	23.9	−0.04
Sex assigned at birth (female or male)	0.45	0.67	−0.31	0.58	−0.17
BMI, kg/m^2	23.4	25.0	−0.24	24.0	−0.09
Systolic blood pressure, mm Hg	119.0	114.6	0.33	116.7	0.17
Diastolic blood pressure, mm Hg	74.3	71.2	0.29	71.9	0.22
Hematocrit, %	42.5	42.6	−0.03	42.7	−0.04
Hemoglobin, /µL	14.1	14.5	−0.19	14.4	−0.13
Lymphocytes, /µL	1,944	1,986	−0.05	1,924	0.03
Mean corpuscular volume, fL/red cell	88.2	89.2	−0.14	89.1	−0.13
Neutrophils, /µL	3,501	3,569	−0.04	3,411	0.05
Platelets, /nL	270.0	244.9	0.28	247.9	0.25

before and after matching; after matching, two groups are more balanced in baseline characteristics, with the standardized mean differences of most variables below 0.20, or one-fifth of one pooled standard deviation, which is typically considered good covariate balance [33]. Before proceeding with randomization inference, we conducted a formal diagnostic test for the paired randomization assumption. Specifically, we used Gagnon-Bartsch and Shem-Tov's [34] classification permutation test, a flexible, machine-learning-based, yet exact diagnostic test. The test was implemented using the R package cpt with logistic regression as the classifier. The paired randomization assumption cannot be rejected for the matched cohort data (p-value=0.243); see Figure 7 for the null distribution of the permutation test and the observed test statistic. Hence, we proceeded with randomization inference as our primary analysis. In the event when the randomization assumption is rejected, we recommend proceeding with a biased randomization assumption as primary analysis using the residual sensitivity value method developed in Chen et al. [35].

Figure 8(a) plots the response magnitude among matched study participants. Both regimens elicited strong binding antibody response to gp41 among participants. We first conducted inference under a randomization assumption: two study participants in each matched pair have the same probability of receiving HVTN 086 T1 or HVTN 205 T4; in other words, the matched observational study succeeded in embedding data into a finely stratified randomized experiment [9, 35]. We conducted randomization inference for ITE quantiles of HVTN 086 T1 vs. HVTN 205 T4 using Chen and Li's method (2024) reviewed in Section 3.2 with the stratified Wilcoxon rank sum statistic [21]. Figure 8(b) shows the simultaneous one-sided confidence intervals for selected ITEs $\tau_{(k)}$'s. According to Figure 8(b), the 95 % lower confidence limit for ITE at rank 51 barely exceeds 0. This implies that,

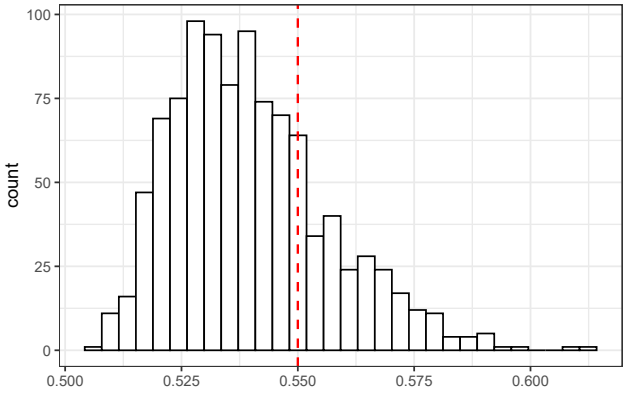
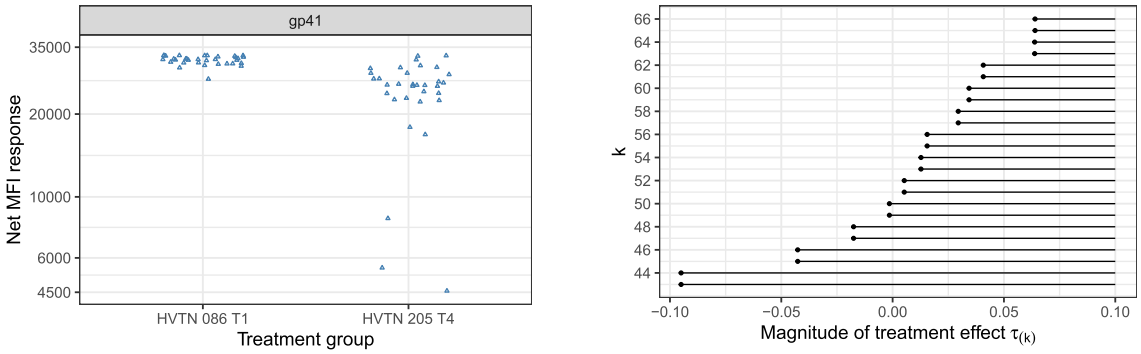


Figure 7: Null distribution of the test statistic in the Classification Permutaiton Test. The red dashed line represents the value of the observed test statistic.



(a) Distributions of observed immune marker measurements

(b) Randomization inference for ITE quantiles

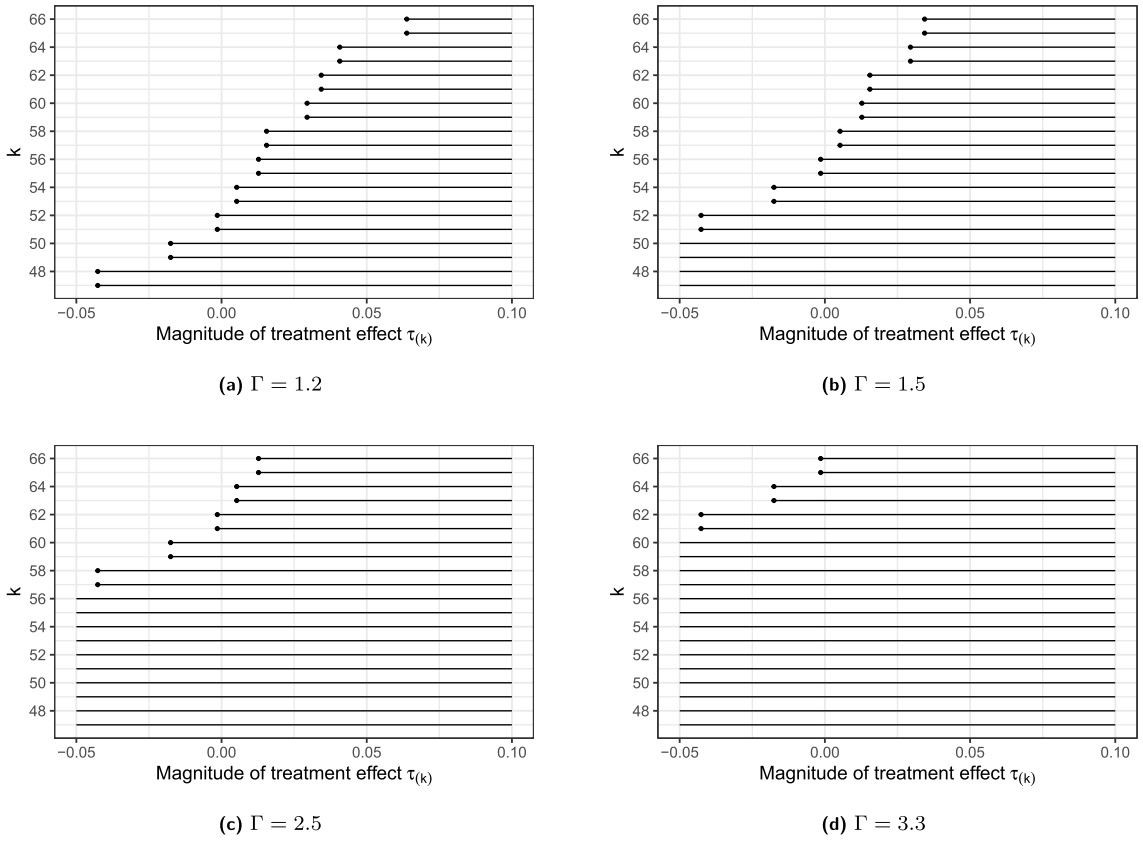
Figure 8: Pooled analysis of HVTN 086 T1 and HVTN 205 T4. (a) Observed serum IgG BAMA responses to gp41 among study participants who received regimen T1 in HVTN 086 and those who received regimen T4 in HVTN 205. (b) 95 % simultaneous one-sided confidence intervals for ITE quantiles of HVTN 086 T1 vs. HVTN 205 T4 using the stratified Wilcoxon rank sum statistic, assuming that there is no hidden confounding.

at 95 % confidence level, at least 24 %, or equivalently 16 out of 66, study participants benefited more from HVTN 086 T1 compared to HVTN 205 T4 in this matched study. We further complemented the analysis by making randomization-based inference for SATE in a matched-pair design [36]. The 95 % lower confidence limit of SATE is 0.086 (in the $\log_{10}$-scale) and quite sensitive due to a few outliers in the HVTN 205 T4 arm. The exact inference for the proportion of study participants who benefited more from HVTN 086 T1 vs. HVTN 205 T4 helps complement a standard analysis of the sample average treatment effect.

A sensitivity analysis assessing deviation from randomization

As is true for all non-randomized studies, the head-to-head comparison between HVTN 086 T1 and HVTN 205 T4 in Section 6.4 could suffer from unmeasured confounding bias. For instance, this would be the case if HVTN 205 enrolled healthier and more immunopotent participants compared to HVTN 086; hence, any treatment effect comparing HVTN 086 T1 to HVTN 205 T4 could be attributed to this hidden bias [9]. We next conduct a sensitivity analysis that investigates how a deviation from the randomization assumption could impact the inferred ITE quantiles in Figure 8(b).

Figure 9 shows the sensitivity analysis results derived from Chen and Li's method (2024) using the stratified Wilcoxon rank sum statistic. Figure 9 plots the 95 % confidence intervals for ITE quantiles under Rosenbaum's sensitivity analysis model indexed by Γ; see Equation (8) [9]. From Figure 9, when the bias in the treatment assignment is as large as Γ=1.5, the ITE at rank 57, i.e., approximately 86 % quantile, remains positive at significance level 0.05, which implies that HVTN 086 T1 would still induce a higher MFI response than HVTN 205 T4 for at least 15 % of study participants in the study under a moderate bias of magnitude 1.5. According to the method described in Rosenbaum and Silber [37], an unobserved covariate associated with at least a 2.5−fold increase in the odds of selecting in the study arm T1/C1 (HVTN 086) as opposed to T4/C4 (HVTN 205) *and* a 2.75−fold increase in the odds of a positive matched-pair difference in MFI values is needed in order to explain away the top 15 % ITE. Such a confounding factor appears unlikely after we have controlled for observed covariates using matching. Therefore, we may conclude that HVTN 086 T1 induces higher MFI response compared to HVTN 205 T4 for at least 15 % of the cohort even in the presence of a moderately large selection. When Γ increases to 3.3, the resulting 95 % confidence intervals no longer cover zero for any participant, implying that we do not have evidence that any participant would benefit more from HVTN 086 T1 vs. HVTN 205 T4 if the trial selection bias is as large as Γ=3.3.



(a) $\Gamma = 1.2$ **(b)** $\Gamma = 1.5$

(c) $\Gamma = 2.5$ **(d)** $\Gamma = 3.3$

Figure 9: The 95 % simultaneous one-sided confidence intervals for ITE quantiles under various sensitivity models indexed by different Γ using the enhanced method in Chen and Li [19].

Discussion

Many recent research has centered around moving beyond Fisher's sharp null hypothesis or Neyman's weak null hypothesis and exploring aspects of the collection of individual treatment effects other than the sample average treatment effect. In this article, we provide a systematic review of relevant methods and closely examine the usefulness of different methods under a wide range of scenarios. We found that randomization inference that tests quantile treatment effects could be a useful complement to the SATE, especially when the scientific interest lies in uncovering and quantifying the heterogeneous treatment effects. In addition, these methods hold promise in helping relevant stakeholders advance an experimental therapy that has a meaningfully large treatment effect on possibly a fraction of study participants, as opposed to a competing therapy that has a similar SATE but nevertheless does not show any treatment effect at a magnitude of practical relevance. Another interesting and practically relevant finding is that, perhaps contrary to expectation, when constructing a simultaneous confidence region jointly for many quantiles, or even every quantile, the cost of multiple hypothesis testing could be minimal.

Overall, our assessment is that the suite of randomization-based inferential methods, testing Fisher's sharp null of no effect whatsoever, Neyman's weak null of no SATE, and various quantiles of ITEs discussed in the current article, should always have a place in empirical researchers' toolbox when analyzing data derived from clinical trials because these methods are always reliable and could potentially yield informative results.

In many senses, quantiles of ITEs are the most fine-grained estimands. This important line of research could be furthered in at least two directions. First, choosing an appropriate test statistic is a key component. As demonstrated in simulation studies, the power of the Stephenson rank sum statistic depends critically on the choice of

s and an optimal choice of the test statistic could depend largely on the data-generating process. One possibility is to develop a data-driven, adaptive approach that produces an optimal or near-optimal choice of the test statistic (or the tuning parameter in the test statistic) or combines several test statistics. Second, we have reviewed and considered methods that (i) assume a constant $Y(0)$; (ii) do not place any restriction on any aspect of the potential outcomes; (iii) place a uniform upper or lower bound on potential outcomes. There are still plenty of other possibilities to leverage auxiliary information from historical data and improve the power of statistical inference. Future work may also explore how to leverage these auxiliary information in experimental designs other than CRE.

Acknowledgment: We would like to thank the AE and one reviewer for constructive feedback.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors declare no competing interest.
Research funding: Research reported in this publication was partially supported by the National Science Foundation (award DMS-2400961 to Xinran Li).
Data availability: The data that support the findings of this study are available from the public-facing HIV Vaccine Trials Network (HVTN) website (https://atlas.scharp.org/). R scripts to reproduce the analyses in this paper are available on the first author's GitHub page (https://github.com/Zhe-Chen-1999/ITE-Inference).

Appendix

```
R function method_SM
## Description
# The function generates one—sided confidence intervals for quantiles of individual
treatment effects in placebo—controlled trials with highly specific endpoints.
## Usage
method_SM(Z, Y, LOD, k_vec, simul = FALSE, nperm = 10^6, Z.perm = NULL, alpha = 0.05,
tol = 0.001)
## Arguments
# Z: An N dimensional treatment assignment vector.
# Y: An N dimensional observed outcome vector.
# LOD: An N dimensional vector specifying the limit of detection of outcomes.
# k.vec: A vector whose coordinates are integers between 1 and N specifying which
quantiles of individual treatment effects is of interest.
# nperm: A positive integer representing the number of permutations for approximating
the randomization distribution of the rank sum statistic.
# Z.perm: A N—by—nperm matrix that specifies the permutated assignments for
approximating the null distribution of the test statistic.
# alpha: A numeric object specifies the level of the confidence interval. In particular,
the confidence level is 1—\alpha.
# tol: A numerical object specifying the precision of the obtained confidence intervals.
For example, if tol = 0.001, then the confidence limits are precise up to 3 digits.
# simul: A logical object indicating whether performing simultaneous inference or
pointwise inference.
## Value
# One—sided upper confidence intervals for specified effect quantiles of interest.
## Examples
```

```
n = 200
m = n * 0.5
Z = sample(c(rep(1, m), rep(0, n-m) ) )
Y = rnorm(n) + Z * rnorm(n, mean = 0, sd = 5)
method_SM(Z = Z, Y = Y, LOD = rep(2,n), k_vec = 1:n, simul = TRUE)
```

References

1. Ditse Z, Mkhize NN, Yin M, Keefer M, Montefiori DC, Tomaras GD, et al. Effect of HIV envelope vaccination on the subsequent antibody response to HIV infection. Msphere 2020;5:e00738–19. https://doi.org/10.1128/msphere.00738-19.
2. Huang Y, Zhang Y, Seaton KE, De Rosa S, Heptinstall J, Carpp LN, et al. Baseline host determinants of robust human HIV-1 vaccine-induced immune responses: a meta-analysis of 26 vaccine regimens. Ebiomedicine 2022;84:104271. https://doi.org/10.1016/j.ebiom.2022.104271.
3. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc 2005;100:322–31.
4. Caughey D, Dafoe A, Li X, Miratrix L. Randomization inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects. J Roy Stat Soc B Stat Methodol 2023, in press.
5. Lipkovich I, Svensson D, Ratitch B, Dmitrienko A. Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. Clin Trials 2023:17407745231174544.
6. Neyman JS. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Ann Agric Sci 1923;10:1–51.
7. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688.
8. Fisher RA. The design of experiments. London and Edinburgh: Oliver and Boyd; 1935.
9. Rosenbaum PR. Observational studies. New York: Springer; 2002.
10. Ding P, Dasgupta T. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. Biometrika 2018;105:45–56.
11. Wu J, Ding P. Randomization tests for weak null hypotheses in randomized experiments. J Am Stat Assoc 2021;116:1898–913.
12. Cohen PL, Fogarty CB. Gaussian prepivoting for finite population causal inference. J Roy Stat Soc B Stat Methodol 2022;84:295–320.
13. Firpo S. Efficient semiparametric estimation of quantile treatment effects. Econometrica 2007;75:259–76.
14. Frölich M, Melly B. Unconditional quantile treatment effects under endogeneity. J Bus Econ Stat 2013;31:346–57.
15. Powell D. Quantile treatment effects in the presence of covariates. Rev Econ Stat 2020;102:994–1005.
16. Fan Y, Park SS. Sharp bounds on the distribution of treatment effects and their statistical inference. Econom Theor 2010;26:931–51.
17. Fan Y, Park SS. Confidence intervals for the quantile of treatment effects in randomized experiments. J Econom 2012;167:330–44.
18. Huang EJ, Fang EX, Hanley DF, Rosenblum M. Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data. Biometrics 2019;75:1228–39.
19. Chen Z, Li X. Enhanced inference for distributions and quantiles of individual treatment effects in various experiments; 2024. Available from: https://arxiv.org/abs/2407.13261.
20. Berger RL, Boos DD. P values maximized over a confidence set for the nuisance parameter. J Am Stat Assoc 1994;89:1012–16.
21. Su Y, Li X. Treatment effect quantiles in stratified randomized experiments and matched observational studies. Biometrika 2023:asad030.
22. Rosenbaum PR. Design of observational studies. New York: Springer; 2010, 10.
23. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J Roy Stat Soc A Stat Soc 2011;174:369–86.
24. Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics 2019;75:685–94.
25. Rosenbaum P. Observation and experiment: an introduction to causal inference. Cambridge, MA: Harvard University Press; 2017.
26. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika 1987;74:13–26.
27. Fogarty CB. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. J Am Stat Assoc 2020;115:1518–30.
28. Sedransk J, Meyer J. Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling. J Roy Stat Soc B (Methodol) 1978;40:239–52.
29. Wang W. Exact optimal confidence intervals for hypergeometric parameters. J Am Stat Assoc 2015, in press. https://doi.org/10.1080/01621459.2014.966191.
30. Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment–control experiments. Proc Natl Acad Sci USA 2018;115:9157–62.

31. Goepfert PA, Elizaga ML, Seaton K, Tomaras GD, Montefiori DC, Sato A, et al. Specificity and six-month durability of immune responses induced by DNA and recombinant modified vaccinia Ankara vaccines expressing HIV-1 virus-like particles. J Infect Dis 2014;210:99—110.
32. Zhang B, Small DS, Lasater KB, McHugh M, Silber JH, Rosenbaum PR. Matching one sample according to two criteria in observational studies. J Am Stat Assoc 2023;118:1140—51.
33. Silber JH, Rosenbaum PR, Trudeau ME, Even-Shoshan O, Chen W, Zhang X, et al. Multivariate matching and bias reduction in the surgical outcomes study. Med Care 2001;39:1048—64.
34. Gagnon-Bartsch J, Shem-Tov Y. The classification permutation test. Ann Appl Stat 2019;13:1464—83.
35. Chen K, Heng S, Long Q, Zhang B. Testing biased randomization assumptions and quantifying imperfect matching and residual confounding in matched observational studies. J Comput Graph Stat 2023;32:528—38.
36. Imai K. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. Stat Med 2008;27:4857—73.
37. Rosenbaum PR, Silber JH. Amplification of sensitivity analysis in matched observational studies. J Am Stat Assoc 2009;104:1398—405.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/scid-2024-0001).