

https://dspace.org/

Generalizing the IntentiontoTreat Effect of an Active Control from Historical PlaceboControlled Trials A Case Study of the Efficacy of Daily Oral TDFFTC in the HPTN 084 Study

Qijia He; F Gao; Ann Delany-Moretlwe; B Zhang

AMER STATISTICAL ASSOC
https://hdl.handle.net/10539/43406
WOS

Downloaded from a DSpace repository. https://dspace.org/

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2025 January 02.

Published in final edited form as:

J Am Stat Assoc. 2024 ; 119(548): 2478–2492. doi:10.1080/01621459.2024.2360643.

Generalizing the intention-to-treat effect of an active control from historical placebo-controlled trials:

A case study of the efficacy of daily oral TDF/FTC in the HPTN 084 study

Qijia He[1], **Fei Gao**[2], **Oliver Dukes**[3], **Sinead Delany-Moretlwe**[4], **Bo Zhang**[*,2]

[1]Department of Statistics, University of Washington

[2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center

[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University

[4]Wits Reproductive Health and HIV Institute, University of the Witwatersrand, Johannesburg, South Africa.

Abstract

In many clinical settings, an active-controlled trial design (e.g., a non-inferiority or superiority design) is often used to compare an experimental medicine to an active control (e.g., an FDA-approved, standard therapy). One prominent example is a recent phase 3 efficacy trial, HIV Prevention Trials Network Study 084 (HPTN 084), comparing long-acting cabotegravir, a new HIV pre-exposure prophylaxis (PrEP) agent, to the FDA-approved daily oral tenofovir disoproxil fumarate plus emtricitabine (TDF/FTC) in a population of heterosexual women in 7 African countries. One key complication of interpreting study results in an active-controlled trial like HPTN 084 is that the placebo arm is not present and the efficacy of the active control (and hence the experimental drug) compared to the placebo can only be inferred by leveraging other data sources. In this article, we study statistical inference for the intention-to-treat (ITT) effect of the active control using relevant historical placebo-controlled trials data under the potential outcomes (PO) framework. We highlight the role of adherence and unmeasured confounding, discuss in detail identification assumptions and two modes of inference (point versus partial identification), propose estimators under identification assumptions permitting point identification, and lay out sensitivity analyses needed to relax identification assumptions. We applied our framework to estimating the intention-to-treat effect of daily oral TDF/FTC versus placebo in HPTN 084 using data from an earlier Phase 3, placebo-controlled trial of daily oral TDF/FTC (Partners PrEP).

Keywords

Active-controlled trial; Compliance; Generalizability; HIV prevention; Intention-to-treat effect; Post-randomization event

[*]Correspondence to Bo Zhang, Assistant Professor of Biostatistics, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, Washington, 98109. bzhang3@fredhutch.org.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1 Introduction

1.1 HIV Prevention Trials Network Study 084: A landmark clinical trial in HIV prevention

The HIV Prevention Trials Network Study 084 (HPTN 084) is a phase 3, double-blind, randomized trial comparing long-acting cabotegravir (CAB-LA), an intramuscular injectable, long-acting form of pre-exposure prophylaxis (PrEP) for HIV prevention, to daily oral tenofovir disoproxil fumarate plus emtricitabine (TDF/FTC) among HIV-uninfected, heterosexual women (Delany-Moretlwe et al., 2022). The study was conducted in 7 countries of sub-Saharan Africa, including Botswana, Eswatini, Kenya, Malawi, South Africa, Uganda, and Zimbabwe. Daily oral TDF/FTC (sold under the brand name Truvada™), a World Health Organization (WHO) recommended PrEP for HIV prevention, has been introduced in these countries; however, despite increasing availability and access to oral PrEP in the region, women have faced considerable barriers, including social stigma, judgement and violence (Delany-Moretlwe et al., 2022), to daily pill-taking, which partly explained why the global HIV prevention efforts have stalled with nearly 1.5 million new HIV infections in 2021, or 4,000 every day, a statistic nearly the same as in 2020. High-risk populations, especially those facing barriers to adhering to the daily oral PrEP, are in urgent need of a long-acting prevention modality like injectible CAB-LA. HPTN 084 reported an HIV incidence of 0.20 per 100 person-years in the CAB-LA arm compared to 1.86 per 100 person-years in the daily TDF/FTC arm (hazard ratio, 0.12; 95% CI, 0.05 to 0.31), demonstrating, *unequivocally*, the superiority of CAB-LA compared to the daily oral TDF/FTC (see Figure S4 in Web Appendix E). Not long after this landmark trial, WHO recommended that “long-acting injectable cabFotegravir (CAB-LA) be offered as an additional HIV prevention option for people at substantial risk of HIV infection” (World Health Organization, 2022).

1.2 Active-controlled trial; intention-to-treat effect; sources of heterogeneity and bias

An important aspect of HPTN 084 is its active-controlled trial design. Active-controlled trials are commonly used in clinical settings to evaluate the safety and effectiveness of an experimental medication compared to a standard therapy (referred to as an active control and abbreviated as AC) when it is unethical to randomize patients to placebo and deprive them of the available standard therapies (Ellenberg and Temple, 2000). Two popular choices of an active-controlled trial design are a superiority design and a non-inferiority (NI) design. In an active-controlled trial design, the placebo arm is not present, so it is not straightforward to estimate the *intention-to-treat* (ITT) effect of the active control compared to the placebo in the trial population (Fleming et al., 2011).

There are two motivations for understanding the ITT effect of an active control compared to the placebo in an active-controlled trial. First, in the design stage of a non-inferiority trial, a key design factor is to select the so-called *NI margin*, defined as an acceptable loss of efficacy comparing the experimental therapy with the AC in the NI trial population. The current standard practice is to set the NI margin to a fraction of the *assumed* ITT effect of the AC; hence, a better understanding of AC’s ITT effect facilitates selecting a rigorous and scientifically justifiable NI margin (Rothmann et al., 2003; James Hung et al., 2003; Fleming et al., 2011). Second, in a post-hoc analysis of the active-controlled trial data, the ITT effect

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of AC versus placebo can be used to establish the ITT effect of the experimental drug versus placebo. The ITT effect of an experimental drug plays a key role in designing future trials to evaluate other experimental drugs, where the current experimental drug may serve as a comparator. In addition, it provides evidence to quantify the experimental drug's public health impact and facilitates comparison of the experimental drug to other therapeutics. Lastly, the ITT effect of the experimental drug helps evaluate how much society should be willing to pay for the improved efficacy of the experimental drug compared to the active control. For instance, Neilan et al. (2022) evaluated the cost-effectiveness of CAB-LA using the Cost-Effectiveness of Preventing AIDS Complications model and a key model parameter in this analysis is the ITT effect of CAB-LA versus placebo. What's more, additional HIV prevention modalities, like an HIV vaccine (Fauci, 2017) and monoclonal antibodies (Miner et al., 2021), are currently under development. The placebo-controlled intention-to-treat effect of CAB-LA serves as an important benchmark to these new interventions.

There are at least three sources of heterogeneity that complicate generalizing an AC's ITT effect from any historical, randomized, placebo-controlled trial to the planned active-controlled trial. First, the actual treatment effect of the AC could be heterogeneous (*treatment effect heterogeneity*). Second, within the same study, different participants could have different probabilities of adhering to the assigned treatment (*within-trial compliance heterogeneity*); for example, in the field of HIV prevention, it was reported that age was correlated with adherence to the prescribed PrEP dose (Grant et al., 2014). Moreover, the same AC could be implemented differently across trials and even the same participants could respond differently to distinct implementations (*between-trial compliance heterogeneity*). Third, trials could target different populations, and therefore, key demographic and health information could differ among trial populations (*target population heterogeneity*). An interplay among treatment effect heterogeneity, within- and between-trial compliance heterogeneity, and target population heterogeneity may lead to generalization bias (Stuart et al., 2011) of the ITT effect. In fact, ITT estimates of the same intervention often differ across historical trials (see, e.g., Table S3 in Web Appendix E). In an editorial discussing discrepancies among these findings, Cohen and Baden (2012) concluded:

> Why the results differ across the various studies reported to date is unclear. However, important considerations include the populations studied; the likely routes of HIV transmission (vaginal vs. anal mucosa)…and most important, medication adherence by study participants.

Cohen and Baden's (2012) comments echo three of the aforementioned sources of heterogeneity.

1.3 Current FDA guidelines; existing approaches and literature; our contributions

Current FDA guidelines for designing an NI trial recommend two strategies for estimating the efficacy of an AC in the planned NI trial from historical evidence (Food and Drug Administration, 2016). First, one may choose a historical placebo-controlled trial of the AC and assume that its ITT effect would remain unchanged in the target NI trial. This assumption is known as the "constancy assumption" (Fleming et al., 2011) and could be implausible considering various sources of heterogeneity previously discussed.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Alternatively, one may employ a meta-analytical approach and derive an *average* estimate based on summary statistics of multiple historical trial results and a random-effects model. The meta-analytical approach acknowledges the variability of ITT estimates across historical trials and incorporates uncertainty quantification using random effects; however, the method is still largely *ad hoc* and is not underpinned by clear *identification* assumptions. Either way, the FDA guidelines recommend acknowledging the unreliability of generalization and using a "discounted" estimate as a means of protection against the generalization bias.

Some authors acknowledge the important role of observed covariates in generalizing the intention-to-treat effect, and have proposed covariate adjustment methods under a "conditional constancy" assumption, that is, the intention-to-treat effect within the same strata of study participants (defined by their observed covariates) is constant across trials (Zhang, 2009). Zhang et al. (2014) developed a sensitivity analysis method that tackles residual inconstancy due to unmeasured confounding. The conditional constancy assumption improves upon the constancy assumption and addresses the target population heterogeneity; however, even the conditional constancy assumption is hard to justify because of the across-trial compliance heterogeneity arising from different implementation strategies. Another unsolved issue concerns unmeasured confounders: What is the precise role of unmeasured confounders in preventing generalization of the ITT effect?

Under the conditional constancy assumption, recent developments in the generalization and transportation methods for causal inference could be directly leveraged to generalize the ITT effect from a historical trial to the planned active-controlled trial (Stuart et al., 2011; Dahabreh et al., 2019); see, e.g., Degtiar and Rose (2021) for a recent review. Pearl (2011, Section 6, Equation 24) discussed identification of the causal effect in the presence of a post-randomization surrogate endpoint under a sequential ignorability assumption (Joffe and Greene, 2009). Rudolph and van der Laan (2017) proposed targeted maximum likelihood estimators (TMLEs) to transport the intention-to-treat effect across populations under a version of the conditional constancy assumption. They consider a different setting where covariates, treatment assignment and treatment received are observed in both reference and trial populations, but there is only follow up data in the reference population. Further, we develop a distinct instrumental variable-based identification strategy that leads to different estimators of the ITT effect. More recently, Dahabreh et al. (2022) discussed in detail unidentifiability of the ITT effect when there are unmeasured common causes of trial participation and treatment, and interpretation of the covariate-standardized ITT estimates (under the conditional constancy assumption) as estimating the effects of joint interventions that scale-up the trial and assign the treatment. In the absence of patient-level data, many authors have proposed meta-analysis-based approaches to estimating causal effects accounting for noncompliance. For instance, Zhou et al. (2019) proposed a Bayesian hierarchical modeling approach to estimating complier average treatment effects and Zhou et al. (2022) proposed a closely related, frequentist approach that targets the same estimand.

In this article, we propose historical-data-driven estimators for AC's ITT effect in a target trial population using relevant historical trials and under different identification assumptions. Our developed framework helps translate, assess, and quantify what FDA guidelines refer to as non-statistically-based uncertainties (Food and Drug Administration, 2016, Page 20)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and places many “essential considerations” raised in Fleming et al. (2011) in the context of formal causal identification assumptions. We assess the finite-sample performance of proposed estimators in simulation studies and apply the proposed estimators to estimating the ITT effect of daily oral TDF/FTC against placebo in the HPTN 084 study using data from this trial and an earlier, historical placebo-controlled trial of daily oral TDF/FTC (Baeten et al., 2012).

2 Notation and framework

2.1 Potential outcomes

We consider the potential outcomes framework (Angrist et al., 1996) to formalize a placebo-controlled trial with noncompliance involving an active control (AC) and a placebo (P). Let $Z_i \in \{0,1\}$ denote a binary treatment assignment (0 for placebo and 1 for AC), and $D_i(Z_i = z_i) \in \{0,1\}$ the potential treatment received had i been assigned the treatment $Z_i = z_i$. Each study participant has a pre-specified probability of receiving either treatment (AC or P). A study participant with $\{D_i(1), D_i(0)\} = \{1,0\}$ complies with the treatment assignment and is referred to as a complier. A participant with $\{D_i(1), D_i(0)\} = \{1,1\}$ is referred to as an always-taker, $\{D_i(1), D_i(0)\} = \{0,0\}$ a never-taker, and $\{D_i(1), D_i(0)\} = \{0,1\}$ a defier (Angrist et al., 1996). We have assumed the Stable Unit Treatment Value Assumption (SUTVA) in the definition of $D_i(Z_i = z_i)$ so that a study participant’s treatment received depends only on the person’s own treatment assignment (Rubin, 1980; Angrist et al., 1996). Each participant is also associated with potential outcomes $\{Y_i(d_i, z_i), d_i \in \{0,1\}, z_i \in \{0,1\}\}$ where we again assume the SUTVA in this definition. Under the exclusion restriction assumption, we further have $Y_i(d_i, z_i) = Y_i(d_i)$, that is, the treatment assignment affects the outcome only via the actual treatment received. Next, we assume Z is randomly assigned and is “relevant” in the sense that $\mathbb{E}[D(Z = 1) - D(Z = 0)] \neq 0$. The SUTVA, exclusion restriction, relevance, and random assignment will be referred to as “core IV assumptions” in this article. Additional assumptions like “monotonicity” ($D_i(Z_i = 1) \geq D_i(Z_i = 0)$) and “one-sided noncompliance” ($D_i(Z_i = 0) = 0$) help further simplify potential outcomes; however, we do not *a priori* make these additional assumptions, though we will consider these as important special cases.

We use the indicator S to denote the trial membership of a participant: $S = t$ if a participant is in the target active-controlled trial; $S = h$ if a participant is in a generic historical placebo-controlled trial. In the later development, we will also explore scenarios where data from two historical trials may be leveraged; in this case, we will use h_1 and h_2 to distinguish distinct historical trials. Regardless of trial membership, each participant is associated with a vector of baseline covariates $\mathbf{X}$. We will use $\mathscr{P}_t$ to denote the joint distribution of $\mathbf{X}$ in the target trial and $\mathscr{P}_h$ that in the historical trial h. We use $\mathbb{E}_{\mathbf{X} \in \mathscr{P}_t}[\cdot]$ and $\mathbb{E}_{\mathbf{X} \in \mathscr{P}_h}[\cdot]$ to denote taking expectation over $\mathscr{P}_t$ and $\mathscr{P}_h$. Throughout, we will make a positivity assumption that $f(S = h \mid \mathbf{X} = \mathbf{x}) > 0$ for all $\mathbf{x}$ in the support of $\mathbf{X}$ in the planned NI trial.

2.2 Estimands

The conditional ITT effect of AC versus P in the historical trial is defined as $ITT(\mathbf{X}; S = \mathrm{h}) = \mathbb{E}[Y(Z = 1) - Y(Z = 0) \mid \mathbf{X}, S = \mathrm{h}]$. Averaging $ITT(\mathbf{X}; S = \mathrm{h})$ over the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

distribution of observed covariates $\mathbf{X} \in \mathscr{P}_h$ then yields the average intention-to-treat effect in the historical trial $S = h$:

$$ITT(S = \mathrm{h}) = \mathbb{E}_{\mathbf{X} \in \mathscr{P}_h}[\mathbb{E}[Y(Z = 1) - Y(Z = 0) \mid \mathbf{X}, S = \mathrm{h}]], \tag{1}$$

which was unbiasedly estimated in the historical placebo-controlled trial by virtue of randomization.

In parallel, we use $ITT(\mathbf{X}; S = \mathrm{t}) = \mathbb{E}[Y(Z = 1) - Y(Z = 0) \mid \mathbf{X}, S = \mathrm{t}]$ to denote the conditional ITT effect of AC versus P in a hypothetical placebo-controlled trial in the target trial population. The intention-to-treat effect of AC in the target trial population is then obtained by averaging $ITT(\mathbf{X}; S = \mathrm{t})$ over the target AC trial population as follows:

$$ITT(S = \mathrm{t}) = \mathbb{E}_{\mathbf{X} \in \mathscr{P}_t}[\mathbb{E}[Y(Z = 1) - Y(Z = 0) \mid \mathbf{X}, S = \mathrm{t}]]. \tag{2}$$

As discussed in Section 1.2, the NI margin and the ITT effect of the experimental drug can be immediately determined once $ITT(S = \mathrm{t})$ is determined. The causal parameter $ITT(S = \mathrm{t})$ is of primary scientific interest and hence our *target parameter*.

2.3 The constancy assumption

The constancy assumption in the NI trial literature (Fleming et al., 2011) states the following relationship between the ITT effect of AC in a target NI trial and that in a chosen historical trial:

Assumption 1 (Constancy). *Let* $ITT(S = t)$ *and* $ITT(S = h)$ *be defined as in* (2) *and* (1), respectively. The constancy assumption is said to hold if $ITT(S = t) = ITT(S = h)$.

Another version of the constancy assumption, referred to as the *conditional constancy assumption* (Zhang, 2009), states the following:

Assumption 2 (Conditional constancy). *Let* $\mathbf{X}$ denote a vector of observed covariates collected in the planned NI trial. Let $ITT(\mathbf{X}; S = t)$ and $ITT(\mathbf{X}; S = h)$ denote conditional intention-to-treat effects in the planned NI trial and the chosen historical trial, respectively. Then the conditional constancy assumption is said to hold if $ITT(\mathbf{X}; S = t) = ITT(\mathbf{X}; S = h)$.

It is transparent from definitions (1) and (2) that when trials enroll study participants from different populations, that is, when $\mathscr{P}_t \neq \mathscr{P}_h$, then Assumption 1 could fail even when Assumption 2 holds. This has been discussed in great detail in the context of *generalizability and transportability* by many authors; see, e.g., Dahabreh et al. (2019). Under Assumption 2, the ITT effect in the target active-controlled trial is identified from the observed data of the historical trial $S = h$ and may be estimated using outcome regression, inverse-probability-weighting, or a doubly-robust combination of both; see, e.g., Dahabreh et al. (2019, Section 5). Although weaker than Assumption 1, Assumption 2 is still a strong assumption; it is,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

after all, a statement about the *intention-to-treat* effect, not the actual treatment effect. Even when interventions have consistent treatment effects for similar study participants across different trials, interventions may be implemented differently, induce different compliance even among similar study participants, and lead to different ITT effects. This is particularly true in HIV prevention studies with daily oral PrEP (Cohen and Baden, 2012).

3 Identification assumptions; a road map for estimation and sensitivity analysis

In this section, we replace the constancy assumption with a set of assumptions regarding effect homogeneity and generalizability. These assumptions are not necessarily weaker; however, they are transparent, problem-specific, and more amenable to being assessed and critiqued. They also motivate the estimation procedures and associated sensitivity analyses.

3.1 No-interaction/homogeneity-type assumption

Intuitively, a statement about the intention-to-treat effect implicitly entails a statement about the compliance structure and the actual treatment effect. Unlike Z, the actual treatment received D is a post-randomization event and *not* randomized. Some version of a homogeneity or no-interaction assumption is therefore necessary to link the ITT effect to the average treatment effect (Swanson et al., 2018, Section 5). Below, we adopt one version from Wang and Tchetgen Tchetgen (2018).

Assumption 3 (No-interaction). *Let* U *denote unmeasured covariates that confound* D*'s effect on* Y*. The no-interaction assumption holds if there is no additive* $U - D$ *interaction in* $\mathbb{E}[Y(D) \mid \mathbf{X}, U]$:

$$\mathbb{E}[Y(D=1) \mid \mathbf{X}, U] - \mathbb{E}[Y(D=0) \mid \mathbf{X}, U] = \mathbb{E}[Y(D=1) \mid \mathbf{X}] - \mathbb{E}[Y(D=0) \mid \mathbf{X}]. \quad \text{(Assumption 3a)}$$

or *no additive* $U - Z$ *interaction in* $\mathbb{E}[D(Z) \mid \mathbf{X}, U]$:

$$\mathbb{E}[D(Z=1) \mid \mathbf{X}, U] - \mathbb{E}[D(Z=0) \mid \mathbf{X}, U] = \mathbb{E}[D(Z=1) \mid \mathbf{X}] - \mathbb{E}[D(Z=0) \mid \mathbf{X}] \quad \text{(Assumption 3b)}$$

Assumption 3 holds if either Assumption 3a or Assumption 3b holds. Assumption 3a holds if there are no more modifiers of D's effect on Y beyond those captured by $\mathbf{X}$. Assumption 3a does *not* hold, for instance, if some genetic factor is suspected to modify D's effect on Y. Fleming et al. (2011) describe an example where the effect of epidermal growth factor receptor-inhibiting drugs in colorectal cancer patients depends strongly on whether tumors express the wild type or the mutated version of the KRAS gene. In this example, the KRAS gene (U) modifies the effect of drug (D) on colorectal cancer (Y), and Assumption 3a *fails* in an analysis not accounting for it.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Assumption 3 also holds if Assumption 3b holds, that is, when the unmeasured modifier of D's effect on Y does not interact with the treatment assignment Z in predicting the treatment received. In the colorectal cancer example, Assumption 3 would still hold if the KRAS gene does *not* interact with a colorectal cancer patient's treatment assignment in predicting whether or not the patient adheres to the prescribed treatment conditional on **X** (possibly including some easier-to-measure aspects of the tumor). This appears to be more reasonable, at least in some applications.

Assumption 3 is a generic assumption that could be applied to either the target AC trial $S = t$ or a historical trial $S = h$. Assumption 3, when applied to the hypothetical placebo-controlled trial in the planned active-controlled trial population, implies the following decomposition:

$$ITT\left(\mathbf{X}; S = t\right) = \underbrace{\mathbb{E}[Y(D = 1) - Y(D = 0) \mid \mathbf{X}, S = t]}_{CATE(\mathbf{X}; S = t)} \times \underbrace{\mathbb{E}[D(Z = 1) - D(Z = 0) \mid \mathbf{X}, S = t]}_{CC(\mathbf{X}; S = t)}, \tag{3}$$

where the term $CATE(\mathbf{X}; S = t)$ describes the average treatment effect of AC versus P conditional on a study participant's covariates, and the conditional compliance term $CC(\mathbf{X}; S = t)$ describes the effect of treatment assignment on treatment received conditional on a study participant's covariates, both in the planned active-controlled trial.

3.2 Conditional average treatment effect; mean generalizability

To link the active-controlled trial to historical trials, we make the following mean generalizability (also known as mean exchangeability) assumption (Stuart et al., 2011; Dahabreh et al., 2019):

Assumption 4 (Mean generalizability/exchangeability).
$CATE(\mathbf{X}; S = t) := \mathbb{E}[Y(D = 1) - Y(D = 0) \mid \mathbf{X}, S = t] = \mathbb{E}[Y(D = 1) - Y(D = 0) \mid \mathbf{X}, S = h] := CATE(\mathbf{X}; S = h)$

Assumption 4 essentially says that study participants with the same observed covariates **X** would experience the same average treatment effect of D on Y in the hypothetical trial and the selected historical trial $S = h$. The major difference between Assumption 4 and Assumption 2 is that Assumption 4 is a statement about the *actual treatment effect* rather than the *intention-to-treat effect*. Assumption 4 is in some sense the minimal assumption needed to extend inference from a historical trial to a target population (Dahabreh et al., 2019, Section 3).

Assumption 4 can still be violated if there exist multiple versions of an active control or placebo across trials (i.e., Rubin's SUTVA is violated); for instance, this could happen if the active control therapy employed in a historical trial is different from that in the planned active-controlled trial due to difference in dosage or ancillary therapies (Fleming et al., 2011; Food and Drug Administration, 2016). For Assumption 4 to hold, researchers should select a historical trial with an active control as similar as possible to that in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the planned active-controlled trial (e.g., both investigating the same medication at the same dose with near-identical ancillary therapies). It is important to note that Assumption 4 only requires the AC therapy itself be identical between trials, not the methods of implementation or dissemination. A sensitivity analysis that models $CATE(\mathbf{X}; S = t)$ as a fraction of $CATE(\mathbf{X}; S = h)$ should be considered when Assumption 4 is suspected not to hold.

Identification of the conditional average treatment effect from a historical placebo-controlled trial, i.e., $CATE(\mathbf{X}; S = h)$, has been discussed extensively in the literature. Two identification strategies are available. First, *point identification* could be achieved by further imposing Assumption 3 on the selected historical trial. Alternatively, $CATE(\mathbf{X}; S = h)$ is *partially identified* under different sets of identification assumptions, including minimal, core IV assumptions. A partial identification interval bounds the range of possible values of the $CATE(\mathbf{X}; h)$ that are consistent with the observed data. Unlike a confidence interval, a partial identification interval would not shrink to a point even when the sample size goes to infinity, as the true parameter may take a range of values and cannot be point identified; see, e.g., Swanson et al. (2018) for a recent review.

An assumption related to Assumption 4 is given in Rudolph and van der Laan (2017), also in the context of transporting ITT effects under non-compliance: $\mathbb{E}[Y \mid D, Z, \mathbf{X}, S = \mathrm{t}] = \mathbb{E}[Y \mid D, Z, \mathbf{X}, S = h]$. This is arguably more difficult to interpret than Assumption 4, since it concerns the generalizability of *associations* rather than causal effects. In our analysis, we explicitly assume that it is the CATE that generalizes, which we identify by leveraging randomization as an instrument.

3.3 Conditional compliance

The conditional compliance term is a difference between $CC_{AC}(\mathbf{X}; S = t) := \mathbb{E}[D(Z = 1) \mid \mathbf{X}, S = t]$ and $CC_P(\mathbf{X}; S = t) := \mathbb{E}[D(Z = 0) \mid \mathbf{X}, S = t]$. It then suffices to identify each term separately. The former term equals $\mathbb{E}[D \mid \mathbf{X}, S = t, Z = 1]$ and is identified based on the compliance data from the active-controlled trial by virtue of randomization. The latter term $CC_P(\mathbf{X}; S = t)$ is not identified from the active-controlled trial, but may be estimated using relevant historical trial data under the following placebo-arm compliance generalizability assumption:

Assumption 5 (Placebo-arm compliance generalizability/exchangeability). $\mathbb{E}[D(Z = 0) \mid \mathbf{X}, S = t] = \mathbb{E}[D(Z = 0) \mid \mathbf{X}, S = h]$.

Note that $\mathbb{E}[D(Z = 0) \mid \mathbf{X}, S = h] = \mathbb{E}[D \mid \mathbf{X}, S = h, Z = 0]$ is directly estimable from historical trial data by randomization. Alternatively, researchers may do a sensitivity analysis for $CC_P(\mathbf{X})$ by varying it from 0 to a sensible value. If the active control therapy is not available in the AC trial population, then one would reasonably set $CC_P(\mathbf{X}) = 0$. Researchers may also vary $CC_P(\mathbf{X})$ in a sensitivity interval centered around $\mathbb{E}[D \mid \mathbf{X}, S = h, Z = 0]$. Either way, instead of outputting a point estimate of $CC_P(\mathbf{X})$ and $CC(\mathbf{X})$, one may output a plausible range.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In applications where the intention-to-treat effect of AC is an important design factor, e.g., when selecting the NI margin in the design phase of the NI trial, researchers do not have compliance data from the NI trial and need to identify the entire conditional compliance term using data from a historical trial under the following mean compliance generalizability assumption:

Assumption 6 (Mean compliance generalizability/exchangeability).
$\mathbb{E}[D(Z=1)-D(Z=0)\mid \mathbf{X}, S=t]=\mathbb{E}[D(Z=1)-D(Z=0)\mid \mathbf{X}, S=h]$.

By randomization of Z, the quantity $\mathbb{E}[D(Z=1)-D(Z=0)\mid \mathbf{X}, S=h]$ equals $\mathbb{E}[D\mid \mathbf{X}, S=h, Z=1]-\mathbb{E}[D\mid \mathbf{X}, S=h, Z=0]$ and is directly estimable from historical trial data. For Assumption 6 to hold (or approximately hold), the selected historical trial should have a near-identical implementation strategy of the AC as in the active-controlled trial. Researchers are also advised to relax Assumption 6 in a sensitivity analysis that varies the conditional compliance term in a sensitivity interval around $\mathbb{E}[D(Z=1)-D(Z=0)\mid \mathbf{X}, S=h]$.

3.4 Summary of identification strategies and non-statistically-based uncertainties

Figure 1 summarizes four aspects we have discussed so far: (i) identification assumptions, including those necessary to identify causal quantities in a trial with non-compliance and those necessary to generalize inference across trials, (ii) quantities involved in the estimation procedure, (iii) different modes of identification, including point or partial identification, and identification using historical trials alone or historical data plus partial AC trial data, and (iv) sensitivity analyses relaxing core assumptions. Together, they help quantify what FDA guidelines refer to as "*non-statistically-based uncertainties*" (Food and Drug Administration, 2016, Page 20). We next discuss *statistically-based uncertainties*, that is, those associated with sampling variability, by formally proposing estimators for the target parameter.

4 Estimation and inference

We consider two scenarios for estimation and inference, each corresponding to one major scientific objective of estimating the ITT effect of the AC versus placebo. We first consider the design stage where researchers have access to data from the historical trial $S=h_1, \mathscr{D}_{h_1}=\{(\mathbf{X}_i, Z_i, D_i, Y_i, S_i=h_1): i=1,\ldots,N_1\}$, data from a second historical trial $S=h_2, \mathscr{D}_{h_2}=\{(\mathbf{X}_i, Z_i, D_i, S_i=h_2): i=N_1+1,\ldots,N_1+N_2\}$, and baseline covariates data from the target AC trial $\mathscr{D}_t=\{(\mathbf{X}_i, S_i=t): i=N_1+N_2+1,\ldots,N_1+N_2+N\}$. Researchers will attempt to leverage the historical data in $S=h_1$ to estimate $CATE(\mathbf{X}; S=t)$ and data in $S=h_2$ to estimate $CC(\mathbf{X}; S=t)$. In this scenario, we write $\mathscr{D}=\mathscr{D}_{h_1}\cup\mathscr{D}_{h_2}\cup\mathscr{D}_t$, where $|\mathscr{D}|$ denote its cardinality. We next consider a more stylistic case where the interest lies in estimating the ITT effect of the AC and hence the experimental therapy versus placebo in an *post hoc* analysis after seeing the compliance data from the target AC trial. In this case, researchers would have access to data $\mathscr{D}_{h_1}$ as mentioned previously plus data $\mathscr{D}_t=\{(\mathbf{X}_i, Z_i, D_i, S_i=t): i=N_1+1,\ldots,N_1+N\}$ from the target AC trial. In this second scenario, we write $\mathscr{D}=\mathscr{D}_{h_1}\cup\mathscr{D}_t$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4.1 Estimation and inference in the design stage

We first consider estimating $ITT(S = t)$ in the design stage and derive a regression-based, historicaldata-driven estimator under Assumption 3 (applied to both $S = t$ and $S = h_1$), Assumption 4 and Assumption 6. Under Assumption 3, the conditional average treatment effect in the trial $S = h_1$, i.e., $CATE(\mathbf{X}; h_1)$, is identified as follows:

$$\mathbb{E}[Y(D=1) - Y(D=0) \mid \mathbf{X}, S = h_1] = \frac{\mathbb{E}[Y \mid \mathbf{X}, S = h_1, Z = 1] - \mathbb{E}[Y \mid \mathbf{X}, S = h_1, Z = 0]}{\mathbb{E}[D \mid \mathbf{X}, S = h_1, Z = 1] - \mathbb{E}[D \mid \mathbf{X}, S = h_1, Z = 0]} = \frac{\delta^Y(\mathbf{X}; h_1)}{\delta^D(\mathbf{X}; h_1)}, \tag{4}$$

where

$$\delta^Y(\mathbf{X}; s) = \mathbb{E}[Y \mid \mathbf{X}, S = s, Z = 1] - \mathbb{E}[Y \mid \mathbf{X}, S = s, Z = 0],$$
$$\delta^D(\mathbf{X}; s) = \mathbb{E}[D \mid \mathbf{X}, S = s, Z = 1] - \mathbb{E}[D \mid \mathbf{X}, S = s, Z = 0].$$

The expression (4) is sometimes known as the *conditional Wald estimand*. It also identifies the conditional complier average treatment effect, if we further assume *monotonicity* in the historical trial (Angrist et al., 1996). Suppose that we obtain $\hat{\delta}^Y(\mathbf{X}; h_1)$ and $\hat{\delta}^D(\mathbf{X}; h_1)$ by fitting correctly specified parametric models for $\mathbb{E}[Y \mid \mathbf{X}, S = h_1, Z = z]$ and $\mathbb{E}[D \mid \mathbf{X}, S = h_1, Z = z]$ and that these models are indexed by finite-dimensional parameters which are estimated, for instance, via maximum likelihood. Then a regression-based estimator $\widehat{CATE}(\mathbf{X}; h_1)$ is obtained as $\widehat{CATE}(\mathbf{X}; h_1) = \hat{\delta}^Y(\mathbf{X}; h_1)/\hat{\delta}^D(\mathbf{X}; h_1)$. As discussed by Wang and Tchetgen Tchetgen (2018), a limitation of this approach with a binary outcome is that one may obtain estimates of $CATE(\mathbf{X}; h_1)$ outside of the $[-1,1]$ interval. A regression-based estimator of conditional compliance in the historical trial h_2 can be analogously obtained as $\widehat{CC}(\mathbf{X}; h_2) = \hat{\delta}^D(\mathbf{X}; h_2)$, where $\hat{\delta}^D(\mathbf{X}; h_2)$ denotes an estimator for the unknown $\delta^D(\mathbf{X}; h_2)$ obtained from $\mathscr{D}_{h_2}$ via parametric regression modelling of $\mathbb{E}[D \mid \mathbf{X}, S = h_2, Z = z]$. By averaging $\widehat{CATE}(\mathbf{X}; h_1)$ and $\widehat{CC}(\mathbf{X}; h_2)$ over $\mathbf{X} \in \mathscr{D}_t$, we obtain the following regression-based estimator of $ITT(S = t)$:

$$\widehat{ITT}_{\text{full, reg}} = \frac{1}{|\mathscr{D}_t|}\sum_{i=1}^{|\mathscr{D}|} \mathbb{1}\{S_i = t\} \times \left\{\widehat{CATE}(\mathbf{X}_i; h_1) \times \widehat{CC}(\mathbf{X}_i; h_2)\right\} = \frac{1}{|\mathscr{D}_t|}\sum_{i=1}^{|\mathscr{D}|} \mathbb{1}\{S_i = t\} \times \frac{\hat{\delta}^Y(\mathbf{X}_i; h_1)}{\hat{\delta}^D(\mathbf{X}_i; h_1)} \hat{\delta}^D(\mathbf{X}_i; h_2). \tag{5}$$

By standard M-estimation theory (Stefanski and Boos, 2002), the estimator $\widehat{ITT}_{\text{full, reg}}$ is a consistent and asymptotically normal estimator for the target parameter $ITT(S = t)$ under the modeling assumptions previously discussed. To obtain a confidence interval, one may use an empirical sandwich variance estimator or the non-parametric bootstrap (Cheng and Huang, 2010).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The regression-based estimator $\widehat{ITT}_{\text{full, reg}}$ is expected to perform well if parametric models are correctly specified. Below, we describe an estimator derived from semiparametric efficiency theory (Bickel et al., 1993), which allows for more flexible estimation of nuisance functions using modern statistical learning approaches, whilst still facilitating parametric-rate inference on the target parameter. It is developed from the same general theory as recent developments in *de-biased machine learning* (Chernozhukov et al., 2017) and *targeted learning* (van der Laan and Rose, 2011).

Recall that the target parameter $ITT(S = t)$ can be expressed as the functional

$$\psi = \mathbb{E}\left[\frac{\delta^Y(\mathbf{X}; h_1)}{\delta^D(\mathbf{X}; h_1)}\delta^D(\mathbf{X}; h_2)\middle| S = t\right].$$

Theorem 1 gives our main result on semiparametric inference.

Theorem 1. *Under a non-parametric model $\mathscr{M}$ that places no restrictions on the observed data distribution, the efficient influence function (EIF) for ψ is equal to*

$$EIF_\psi = \frac{1}{\kappa}\frac{(2Z-1)\mathbb{1}(S=h_1)}{f(Z \mid \boldsymbol{X}, S=h_1)}\frac{f(S=t \mid \mathbf{X})}{f(S=h_1 \mid \mathbf{X})}\frac{\delta^D(\mathbf{X}; h_2)}{\delta^D(\mathbf{X}; h_1)}\left[Y - \mu_{Y,0}(\mathbf{X}; h_1) - \{D - \mu_{D,0}(\mathbf{X}; h_1)\}\frac{\delta^Y(\mathbf{X}; h_1)}{\delta^D(\mathbf{X}; h_1)}\right]$$
$$+ \frac{1}{\kappa}\frac{(2Z-1)\mathbb{1}(S=h_2)}{f(Z \mid \boldsymbol{X}, S=h_2)}\frac{f(S=t \mid \mathbf{X})}{f(S=h_2 \mid \mathbf{X})}\frac{\delta^Y(\mathbf{X}; h_1)}{\delta^D(\mathbf{X}; h_1)}\left\{D - \mu_{D,0}(\mathbf{X}; h_2) - \delta^D(\mathbf{X}; h_2)Z\right\} + \frac{1}{\kappa}\mathbb{1}\left(S=t\right)$$
$$\left\{\frac{\delta^Y(\mathbf{X}; h_1)}{\delta^D(\mathbf{X}; h_1)}\delta^D(\mathbf{X}; h_2) - \psi\right\},$$

where $\mu_{Y,z}(\mathbf{X}; s) = \mathbb{E}[Y \mid \mathbf{X}, S = s, Z = z]$, $\mu_{D,z}(\mathbf{X}; s) = \mathbb{E}[D \mid \mathbf{X}, S = s, Z = z]$ and $\kappa = f(S = t)$. The semiparametric efficiency bound under $\mathscr{M}$ is $\mathbb{E}\left[EIF_\psi^2\right]$.

Although $\mathscr{M}$ is a non-parametric model, the treatment assignment probabilities $f(Z = z \mid \mathbf{X}, S)$ are known by design in our setting and in particular do not typically depend on $\mathbf{X}$. Nevertheless, it follows, e.g. from Hahn (1998), that knowledge of $f(Z = z \mid \mathbf{X}, S)$ in this case should not change the bound. In contrast, one may be able to leverage information on $f(S = s \mid \mathbf{X})$ to gain precision, although we do not pursue this since such knowledge may not generally be available.

To construct an estimator of ψ based on EIF_ψ, one must estimate $\delta^Y(\mathbf{X}; h_1)$, $\delta^D(\mathbf{X}; h_1)$ and $\delta^D(\mathbf{X}; h_2)$, plus the additional nuisance functions $\mu_{Y,z}(\mathbf{X}; s)$, $\mu_{D,z}(\mathbf{X}; s)$ and $f(S = s \mid \mathbf{X})$. Although $f(Z \mid \mathbf{X}, S = h_1)$ is known, one typically uses the *estimated* version of it. One strategy would be to develop a multiply robust approach similar to that in Wang and Tchetgen Tchetgen (2018), based on parametric working models for the nuisance functions. Here, we describe an alternative approach, which allows for off-the-shelf methods to learn these quantities. These could include classical non-parametric estimators (e.g. kernel smoothers, sieves) or potentially more flexible statistical learning approaches (random forests, kernel ridge regression, Lasso, ensemble methods). After obtaining estimates

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\hat{\delta}^Y(\mathbf{X};h_1), \hat{\delta}^D(\mathbf{X};h_1), \hat{\delta}^D(\mathbf{X};h_2), \hat{f}\left(Z \mid \mathbf{X}, S=s\right), \hat{f}\left(S=s \mid \mathbf{X}\right), \hat{\mu}_{Y,0}(\mathbf{X};h_1)$ and $\hat{\mu}_{D,0}(\mathbf{X};h_2)$, one can then estimate ψ as

$$\widehat{ITT}_{\text{EIF}} = \frac{1}{|\mathscr{D}_t|}\sum_{i=1}^{|\mathscr{D}|}\frac{(2Z_i-1)\mathbb{1}(S_i=h_1)}{\hat{f}(Z_i \mid \mathbf{X}_i, S_i=h_1)}\frac{\hat{f}(S_i=t \mid \mathbf{X}_i)}{\hat{f}(S_i=h_1 \mid \mathbf{X}_i)}\frac{\hat{\delta}^D(\mathbf{X}_i;h_2)}{\hat{\delta}^D(\mathbf{X}_i;h_1)}\left[Y_i - \hat{\mu}_{Y,0}(\mathbf{X}_i;h_1) - \{D_i - \hat{\mu}_{D,0}(\mathbf{X}_i;h_1)\}\frac{\hat{\delta}^Y(\mathbf{X}_i;h_1)}{\hat{\delta}^D(\mathbf{X}_i;h_1)}\right] + \frac{1}{|\mathscr{D}_t|}\sum_{i=1}^{|\mathscr{D}|}\frac{(2Z_i-1)\mathbb{1}(S_i=h_2)}{\hat{f}(Z_i \mid \mathbf{X}_i, S_i=h_2)}\frac{\hat{f}(S_i=t \mid \mathbf{X}_i)}{\hat{f}(S_i=h_2 \mid \mathbf{X}_i)}\frac{\hat{\delta}^Y(\mathbf{X}_i;h_1)}{\hat{\delta}^D(\mathbf{X}_i;h_1)}\left\{D_i - \hat{\mu}_{D,0}(\mathbf{X}_i;h_2) - \hat{\delta}^D(\mathbf{X}_i;h_2)Z_i\right\} + \frac{1}{|\mathscr{D}_t|}\sum_{i=1}^{|\mathscr{D}|}\mathbb{1}\{S_i=t\}\frac{\hat{\delta}^Y(\mathbf{X}_i;h_1)}{\hat{\delta}^D(\mathbf{X}_i;h_1)}\hat{\delta}^D(\mathbf{X}_i;h_2).$$

Under regularity conditions, if each of the nuisance estimators converges to the truth with mean squared error rate shrinking faster than $n^{-1/4}$ and certain Donkser conditions on the nuisance functions hold (Van der Vaart, 2000), then $\widehat{ITT}_{\text{EIF}}$ is $n^{1/2}$-consistent and asymptotically normal. Furthermore, supposing that $\mathbb{E}[D \mid \mathbf{X}, S=h_1, Z]$ is consistently estimated, the estimator is asymptotically unbiased (although not necessarily $n^{1/2}$-consistent) so long as one of the following restrictions hold: (1) $\mathbb{E}[Y \mid \mathbf{X}, S=h_1, Z]$ and $E[D \mid \mathbf{X}, S=h_2, Z]$ are consistently estimated; (2) $\mathbb{E}[Y \mid \mathbf{X}, S=h_1, Z]$ and $f(S \mid \mathbf{X})$ are consistently estimated; and (3) $\mathbb{E}[D \mid \mathbf{X}, S=h_2, Z]$ and $f(S \mid \mathbf{X})$ are consistently estimated. See Section 4.5 of Wang and Tchetgen Tchetgen (2018) for further discussion about the robustness properties. Additional robustness may be attained by using doubly robust estimators for certain nuisance functions and/or adopting the parametrizations in Wang and Tchetgen Tchetgen (2018). An estimator of the asymptotic variance can be obtained using a sandwich estimator. As discussed in Chernozhukov et al. (2017), if very flexible learning methods are used, sample-splitting (estimating the nuisance functions on a training split, and ψ on a test split) or cross-fitting are recommended to alleviate the Donsker conditions.

4.2 Estimation and inference in the post hoc analysis

The previous results straightforwardly extend to the setting where one wishes to evaluate the ITT effect of the AC versus placebo with data available from the target AC trial. In that case, conditional compliance $\mathbb{E}[D(Z=1) \mid \mathbf{X}, S=t]$ can be identified under randomization as $\mu_{D,1}(\mathbf{X};t) := \mathbb{E}[D \mid \mathbf{X}, S=t, Z=1]$. However, $\mathbb{E}[D(Z=0) \mid \mathbf{X}, S=t]$ cannot be identified as straightforwardly, because there is no placebo arm in the AC trial. We will proceed here, as in our case study, by treating $\mathbb{E}[D(Z=0) \mid \mathbf{X}, S=t]$ as a sensitivity parameter $\mu^*_{D,0}(\mathbf{X};t)$, such that $\delta^{D*}\left(\mathbf{X};t\right) := \mu_{D,1}\left(\mathbf{X};t\right) - \mu^*_{D,0}\left(\mathbf{X};t\right)$ and $\hat{\delta}^{D*}\left(\mathbf{X};t\right) := \hat{\mu}_{D,1}\left(\mathbf{X};t\right) - \mu^*_{D,0}\left(\mathbf{X};t\right)$. In that case, the identification functional is now

$$\psi = \mathbb{E}\left[\frac{\delta^Y(\mathbf{X};h_1)}{\delta^D(\mathbf{X};h_1)}\delta^{D*}\left(\mathbf{X};t\right)\middle| S=t\right].$$

Results on estimation follow closely along the lines described in the previous subsection. Indeed, the regression-based estimators equal

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\widehat{ITT}_{\text{full, reg}} = \frac{1}{|\mathscr{D}_i|}\sum_{i=1}^{|\mathscr{D}|} \mathbb{1}\{S_i = t\} \times \frac{\hat{\delta}^Y(\mathbf{X}_i; h_1)}{\hat{\delta}^D(\mathbf{X}_i; h_1)} \hat{\delta}^{D*}(\mathbf{X}_i; t).$$

whereas the estimators based on the efficient influence function simplify to

$$\widehat{ITT}_{\text{EIF}} = \frac{1}{|\mathscr{D}_i|}\sum_{i=1}^{|\mathscr{D}|} \frac{(2Z_i - 1)\mathbb{1}(S_i = h_1)}{\hat{f}(Z_i \mid \mathbf{X}_i, S_i = h_1)} \frac{\hat{f}(S_i = t \mid \mathbf{X}_i)}{\hat{f}(S_i = h_1 \mid \mathbf{X}_i)} \frac{\hat{\delta}^{D*}(\mathbf{X}_i; t)}{\hat{\delta}^D(\mathbf{X}_i; h_1)}$$
$$\left[Y_i - \hat{\mu}_{Y,0}(\mathbf{X}_i; h_1) - \{D_i - \hat{\mu}_{D,0}(\mathbf{X}_i; h_1)\}\frac{\hat{\delta}^Y(\mathbf{X}_i; h_1)}{\hat{\delta}^D(\mathbf{X}_i; h_1)}\right] + \frac{1}{|\mathscr{D}_i|}\sum_{i=1}^{|\mathscr{D}|} \frac{\mathbb{1}(Z_i = 1)\mathbb{1}(S_i = t)}{\hat{f}(Z_i \mid \mathbf{X}_i, S_i = t)} \frac{\hat{\delta}^Y(\mathbf{X}_i; h_1)}{\hat{\delta}^D(\mathbf{X}_i; h_1)}\{D_i - \hat{\mu}_{D,1}(\mathbf{X}_i; t)\}$$
$$+ \frac{1}{|\mathscr{D}_i|}\sum_{i=1}^{|\mathscr{D}|} \mathbb{1}\{S_i = t\}\frac{\hat{\delta}^Y(\mathbf{X}_i; h_1)}{\hat{\delta}^D(\mathbf{X}_i; h_1)}\hat{\delta}^{D*}(\mathbf{X}_i; t).$$

One can then estimate the ITT effect of the experimental therapy versus placebo in the target AC trial population by adding one of the estimates described above to the estimated ITT comparison of the experimental therapy versus active control. Although the above developments treat $\mu^*_{D,0}(\mathbf{X}_i; t)$ as fixed, in practice, one may wish to very it based on a plausible range of values.

4.3 Extensions

Point identification of $CATE(\mathbf{X}; h_1)$ requires imposing Assumption 3 or other homogeneity-type assumptions on the historical trial $S = h_1$ (Swanson et al., 2018, Section 5.2). Alternatively, one may proceed by constructing partial identification intervals $[L(\mathbf{X}), U(\mathbf{X})]$ such that $CATE(\mathbf{X}; h_1) \in [L(\mathbf{X}), U(\mathbf{X})]$ almost surely. Depending on the assumptions one is willing to make about the treatment assignment and treatment received, different partial identification bounds can be formulated (Swanson et al., 2018). We review some estimation strategies for partial identification bounds in Web Appendix B for completeness. We will construct partial identification bounds that are motivated by our case study in Section 6. Web Appendix C also discusses some variants of the regression-based estimator and a sensitivity analysis assessing Assumption 3.

5 Simulation study

5.1 Goal and structure

We consider data generating processes that have all three three sources of heterogeneity: treatment effect heterogeneity, within- and across-trial compliance heterogeneity, and target population heterogeneity. We generate two historical datasets $\mathscr{D}_{h_1}$ and $\mathscr{D}_{h_2}$, and a hypothetical placebo-controlled trial dataset $\mathscr{D}_{\text{target}}$ according to the following data generating process:

Sample sizes: $N_1 = N_2 = N = 1000,\ 2000$, and 5000.

Observed covariates and overlap: We consider the following two data generating processes for $\mathbf{X}$, one mimicking the case study (Scenario X1) and the other following a standard multivariate normal distribution (Scenario X2).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Scenario X1 : We sample with replacement HPTN 084 participants' observed covariates to form $\mathscr{D}_{\text{target}}$. We then sample Partners PrEP participants' observed covariates to form $\mathscr{D}_{h_1}$ and $\mathscr{D}_{h_2}$ and control the amount of overlap between these two historical datasets and $\mathscr{D}_{target}$ using the following biased sampling strategy. For each study participant in Partners PrEP, we estimate a "probability of trial participation," defined as the probability of selection into the HPTN 084 study over the Partners PrEP study based on a participant's baseline characteristics (Cole and Stuart, 2010; Stuart et al., 2011). This "probability of trial participation" is a version of Rosenbaum and Rubin's (1983) propensity score and captures the covariate balance between the target and historical datasets. By over- and under-sampling participants in Partners PrEP with large estimated "probability of participation," we then control the amount of overlap between datasets. Specifically, the historical dataset $\mathscr{D}_{h_j}$ was formed by sampling $N_{j,\text{high}}$ and $N_{j,\text{low}}$ participants with high (above 0.5) and low (below 0.5) probability of participation, $j = 1,2$. We consider three overlap levels: (i) Poor overlap: $N_{1,\text{high}} = 0.1N_1$, $N_{1,\text{low}} = 0.9N_1$, $N_{2,\text{high}} = 0.15N_2$, $N_{2,\text{low}} = 0.85N_2$; (ii) Limited overlap: $N_{1,\text{high}} = 0.19N_1$, $N_{1,\text{low}} = 0.81N_1$, $N_{2,\text{high}} = 0.19N_2$, $N_{2,\text{low}} = 0.81N_2$; (iii) Sufficient overlap: $N_{1,\text{high}} = 0.4N_1$, $N_{1,\text{low}} = 0.6N_1$, $N_{2,\text{high}} = 0.5N_2$, $N_{2,\text{low}} = 0.5N_2$. To illustrate, Figure 2 plots the overlap between $\mathscr{D}_{\text{target}}$ and $\mathscr{D}_{h_1}$ in poor overlap, limited overlap, and sufficient overlap scenarios.

Scenario X2: We generate a 10-dimensional $\mathbf{X}\sim$ Multivariate Normal $(\boldsymbol{\mu}, 0.5 \cdot \mathbf{Id})$, where $\mathbf{Id}$ is an identity matrix, $\boldsymbol{\mu} = (c, c, c, 0, \ldots, 0)^{\mathrm{T}}$ in $\mathscr{D}_{h_2}$, $\boldsymbol{\mu} = (1.2c, 1.2c, 1.2c, 0, \ldots, 0)^{\mathrm{T}}$ in $\mathscr{D}_{h_1}$, and $\boldsymbol{\mu} = (0.8c, 0.8c, 0.8c, 0, \ldots, 0)^{\mathrm{T}}$ in $\mathscr{D}_{\text{target}}$, and $c \in \{0, 0.25, 0.50\}$. Parameter c controls the amount of overlap in this scenario.

Treatment assignment: Z is Bernoulli (0.5) in $\mathscr{D}_{h_1}, \mathscr{D}_{h_2}$ and $\mathscr{D}_{\text{target}}$.

Treatment received: D is Bernoulli with $P(D(Z) = 1 \mid \mathbf{X}) = \text{expit}\{Z(2.5 - 0.1X_1 + 0.3X_2 - 0.4X_5) + (1 - Z)(-0.3X_1 - 0.4X_5 - 1.5)\}$ in $\mathscr{D}_{h_2}$ and $\mathscr{D}_{\text{target}}$, and expit $\{Z(2 + 0.2X_1 - 0.2X_5) + (1 - Z)(-0.2X_5 - 1)\}$ in $\mathscr{D}_{h_1}$, where expit $(x) = \exp(x)/(1 + \exp(x))$ is the inverse of the logit function.

According to the above data generating process, the covariate distribution $\mathbf{X}$ is distinct among $\mathscr{D}_{h_1}, \mathscr{D}_{h_2}$, and $\mathscr{D}_{\text{target}}$ (that is, target population heterogeneity exists) in Scenario X1 and in Scenario X2 when $c \neq 0$. The effect of Z on D in $\mathscr{D}_{h_1}$ is different from that in $\mathscr{D}_{h_2}$ and $\mathscr{D}_{\text{target}}$ (that is, across-trial compliance heterogeneity exists). Moreover, (X_1, X_2) modify the effect of Z on D in $\mathscr{D}_{h_2}$ and $\mathscr{D}_{\text{target}}$ (that is, within-trial compliance heterogeneity exists) so that the marginal compliance rate is different between $\mathscr{D}_{h_2}$ and $\mathscr{D}_{\text{target}}$.

Outcome: We consider two sets of data generating processes in $\mathscr{D}_{h_1}$ and $\mathscr{D}_{h_2}$: a linear data generating process (Scenario Y1):

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$P\left(Y\left(Z\right)=1\mid\mathbf{X}\right)=$$
$$\begin{cases}\text{expit}\{Z(2.6-0.6X_1-0.8X_2+0.4X_3)+(1-Z)(1.6-0.7X_1-0.7X_2+0.4X_3-0.2X_5)\} & \text{in } \mathscr{D}_{h_1},\\ \text{expit}\{1.4-X_1-0.6X_2+0.4X_3-0.6X_5+3.5Z\} & \text{in } \mathscr{D}_{h_2},\end{cases}$$

and a nonlinear data generating process (Scenario Y2):

$$P\left(Y\left(Z\right)=1\mid\mathbf{X}\right)=$$
$$\begin{cases}\text{expit}\left\{Z\left(2.6-0.6X_1-0.8X_2+0.4\sqrt{X_3}\right)+\left(1-Z\right)\left(1.6-0.7X_1-0.7X_2^3+0.4X_3-0.2X_5\right)\right\} & \text{in } \mathscr{D}_{h_1},\\ \text{expit}\left\{1.4-X_1-0.6\sqrt{X_2}+0.4X_3-0.6X_5+3.5Z\right\} & \text{in } \mathscr{D}_{h_2}.\end{cases}$$

In the hypothetical trial dataset $\mathscr{D}_{\text{target}}$, we generated the potential outcome $P(Y(Z=0)=1\mid\mathbf{X})=0$ and hence $P(Y(Z=1)=1\mid\mathbf{X})$ equals the conditional intention-to-treat effect which is a product of the conditional average treatment effect in $\mathscr{D}_{h_1}$ and the conditional compliance in $\mathscr{D}_{h_2}$. The data-generating process also ensures that $CATE(\mathbf{X};S=h_1)$ is bounded between −1 and 1.

We considered 6 estimators of the ITT effect: (i) a difference-in-means estimator $\widehat{ITT}_{\text{hypo}}$ based on the unobservable outcome data in $\mathscr{D}_{\text{target}}$, (ii) and (iii) two covariate-adjusted estimators that (incorrectly) assume the conditional constancy assumption between $\mathscr{D}_{\text{target}}$ and $\mathscr{D}_{h_1}\left(\widehat{ITT}_{\text{const},1}\right)$ and between $\mathscr{D}_{\text{target}}$ and $\mathscr{D}_{h_2}\left(\widehat{ITT}_{\text{const},2}\right)$, (iv) a historical-data-driven, regression-based estimator $\widehat{ITT}_{\text{reg, par}}$, (v) a historical-data-driven, EIF-based estimator $\widehat{ITT}_{\text{EIF, par}}$ with all nuisance parameters estimated via parametric regression models, and (vi) a historical-data-driven, EIF-based estimator $\widehat{ITT}_{\text{EIF, gam}}$ with all nuisance parameters estimated via generalized additive models (Hastie, 2017). In each setting, we repeat the simulation 1000 times.

5.2 Results

Figure S1 in Web Appendix D compares the sampling distributions of 6 estimators under consideration when the sample sizes are $N_1=N_2=N=2000$, observed covariates are generated according to Scenario X1, and the outcomes are generated according to Scenario Y1. The ground truth intention-to-treat effects are superimposed using red dashed lines. The three historical-datadriven estimators $\widehat{ITT}_{\text{reg, par}}$, $\widehat{ITT}_{\text{EIF, par}}$, and $\widehat{ITT}_{\text{EIF, gam}}$ all closely resemble the ground truth ITTs, though they have larger variances compared to that of the unobtainable, gold-standard estimator $\widehat{ITT}_{\text{hypo}}$. As the overlap between historical and target datasets improves, the variance of each historical-data-driven estimator starts to shrink and the sampling distribution becomes more concentrated around the ground truth. Table 1 summarizes the percentage of bias and coverage of 95% confidence intervals for different sample sizes and overlap levels. We encountered simulated datasets where the estimator $\widehat{ITT}_{\text{EIF, gam}}$ became unstable due to the small weights in the denominator, especially when the covariate overlap is poor. We reported in the table captions the number of times this phenomenon occurred out of 1000 Monte Carlo replications. In these cases, we applied a hard thresholding and let the estimator be $\phi\left(\widehat{ITT}_{\text{EIF, gam}}\right)$ where function

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\phi(x) = 1, \forall x \geq 1, \phi(x) = -1, \forall x \leq -1$, and $\phi(x) = x$ otherwise. The percentage of bias of $\widehat{ITT}_{\text{EIF, gam}}$ reported in Table 1 is based on the truncated version. Similar to the impression delivered by Figure S1, in all cases considered in this simulation study, three historical-data-driven estimators had small to negligible biases. On the other hand, two estimators based on the incorrect conditional constancy assumption ($\widehat{ITT}_{\text{const,1}}$ and $\widehat{ITT}_{\text{const,2}}$) were heavily biased and their confidence intervals' coverage was nowhere close to the nominal level. The bias that persists after adjusting for the observed covariates difference is often observed in empirical studies and referred to as "residual confounding" by Zhang et al. (2014). Our simulation exhibits concrete settings where such residual confounding could emerge. The 95% confidence intervals for all but $\widehat{ITT}_{\text{EIF, gam}}$ were based on nonparametric bootstrap. We found that the bootstrapped 95% CIs of $\widehat{ITT}_{\text{reg, par}}$ and $\widehat{ITT}_{\text{EIF, par}}$ approximately attained their nominal level when sample sizes are as large as 2000 in each dataset. The bootstrapped CIs of $\widehat{ITT}_{\text{EIF, gam}}$ were found to be highly conservative; on the other hand, the 95% CIs obtained based on asymptotic normality and estimated asymptotic variances tended to undercover when the overlap was poor and the sample size was small, but began to achieve nominal coverage level when the overlap was sufficient and sample size was as large as 5000. In the Web Appendix D.2, we report additional simulation results when observed covariates were generated according to Scenario X2 and outcomes were generated according to Scenario Y1 and Scenario Y2. In the additional nonlinear data generating process Scenario Y2, $\widehat{ITT}_{\text{EIF, gam}}$ continued to have negligible bias and good coverage, while $\widehat{ITT}_{\text{reg, par}}$ became biased once the parametric models became misspecified.

6 Case study: Efficacy of daily TDF/FTC in HIV-1 prevention

6.1 Historical placebo-controlled trials of daily oral TDF/FTC

Our goal is to estimate the ITT effect of daily oral TDF/FTC versus placebo and then the ITT effect of CAB-LA in the HPTN 084 trial population. We consider an integrated analysis of the patient-level data from HPTN 084 and a historical placebo-controlled trial of daily oral TDF/FTC using our proposed framework and methods. There are 3 large-scale, multicenter, randomized trials that evaluated daily oral TDF/FTC: Partners PrEP (Baeten et al., 2012), FEM-PrEP (Van Damme et al., 2012), and VOICE (Marrazzo et al., 2015). The FEM-PrRP and VOICE were conducted in the heterosexual women population in multiple African countries, while the Partners PrEP study enrolled HIV-uninfected heterosexual men and women who had a partner living with HIV (i.e., HIV-1-serodiscordant heterosexual couples) from Kenya and Uganda. These three historical studies recorded quite different annualized HIV incidence in the daily oral TDF/FTC arm. The Partners PrEP study reported an incidence of 0.95 per 100 person-years in the TDF/FTC arm among heterosexual women (Baeten et al., 2012, Figure 3). On the other hand, both the FEM-PrEP study (Van Damme et al., 2012, Table 2) and the VOICE study (Marrazzo et al., 2015, Table 3) reported an incidence of 4.7 per 100 person-years in the TDF/FTC arm. The annualized HIV incidence was reported to be 1.86 per 100 person-years in the TDF/FTC arm of the HPTN 084 study. In this integrated analysis, we chose the Partners PrEP study as the historical placebo-controlled trial because the gap between the annualized HIV incidence rate in the TDF/FTC

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

arm between the Partners PrEP study and the HPTN 084 study, while still substantial, was considerably smaller compared to the other two studies.

6.2 Overlap between the HPTN 084 and Partners PrEP trial populations; target population

We first examine the overlap between trial population of the HPTN 084 study and that of the Partners PrEP study. Because the HPTN 084 study enrolled only female participants, we focused on the female participants in the Partners PrEP study. Moreover, as the Partners PrEP study enrolled heterosexual women who had a partner living with HIV, the study would not reveal the ITT effect or the conditional average treatment effect of daily oral TDF/FTC for heterosexual women whose partners did not live with HIV, if a partner's HIV status is an important modifier of the compliance pattern or the treatment effect. Therefore, instead of making inference for the entire HPTN 084 population, we only focused on about one third of the HPTN 084 participants whose partners either living with HIV or having an unknown HIV status as our target population.

The second and third columns in Table 2 summarize the baseline characteristics of study participants in the target population of HPTN 084 and female participants in the Partners PrEP study. Compared to those in Partners PrEP, participants in the HPTN 084 target population were younger (mean age 26.3 versus 33.5) and received more education (46.3% versus 6.3% completing the secondary school). HPTN 084 participants also had higher unemployment rate (75.8% versus 31.8%), higher positivity rates of baseline diagnoses of gonorrhea (6.0% versus 1.2%), chlamydia (16.3% versus 1.1%) and trichomonas (7.8% versus 6.8%), and lower positivity rate of syphilis (2.6% versus 5.8%). To help better summarize and visualize the covariate overlap between two population, Figure S5 in the Web Appendix E exhibits the distributions of the estimated "probability of participation" in the target HPTN 084 population and among female participants in the Partners PrEP study (Cole and Stuart, 2010; Stuart et al., 2011). The plot suggests that there is overlap between two populations across the spectrum of the "probability of participation," although the overlap is limited so covariate adjustment is warranted.

The first column of Table 2 further exhibits the covariate distribution of the entire HPTN 084 study. We found that participants in the target population were similar to the entire HPTN 084 trial population in age, education, unemployment rate and baseline sexually transmitted infections; nevertheless, because partners' HIV status could be an important risk factor, we still restricted our analysis to $n = 1{,}139$ participants whose partners lives with HIV or had an unknown status.

In addition to baseline characteristics of trial participants, self-reported adherence to the daily pill-taking was also different in HPTN 084 and Partners PrEP. Consuming at least 80% of prescribed pills was typically considered "adhering to the drug" in the HIV prevention literature (Murnane et al., 2015). Adopting this definition, 80.5% of daily TDF/FTC recipients adhered to the prescription in the Partners PrEP Study, and this number was 52.4% in the HPTN 084 study. In the HPTN 084 study, measurements of plasma tenofovir concentrations from a prespecified random cohort of 405 study participants were obtained; 812 out of 1, 939 samples (41.0%) had tenofovir concentrations consistent with daily use ($\geq$ 40ng/mL).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The right panel of Figure 3 plots the cumulative incidence curves in the CAB-LA and TDF/FTC arms in the target population, while the left panel plots the cumulative incidence curves in the TDF/FTC and placebo arms among heterosexual women in the Partners PrEP study. In view of the difference in patient composition and adherence pattern, both the constancy and conditional constancy assumptions are likely to be violated. Below, we seek to estimate the *ITT* effect of daily oral TDF/FTC against placebo for the target population based on evidence from the Partners PrEP study using the framework developed in the article.

6.3 Estimating the ITT effect of daily oral TDF/FTC in the target population: two approaches

We estimated the intention-to-treat effect of daily oral TDF/FTC against placebo in reducing HIV-1 incidence in the target population under both point and partial identification frameworks. First, we assume Assumption 3 holds for the Partners PrEP study and the hypothetical placebo-controlled trial of daily oral TDF/FTC in the target population with the observed covariates including age, employment status, education, and four comorbidties including gonorrhea, chlamydia, trichomonas, and syphilis. Under this assumption and Assumption 4, we estimated the average treatment effect of daily oral TDF/FTC against placebo conditional on observed baseline covariates based on the Partners PrEP data. We then estimated the conditional compliance using the observed self-reported adherence data in the TDF/FTC arm in the HPTN 084 study and by treating the probability that a placebo recipient received daily oral TDF/FTC (i.e., $CC_P(\mathbf{X}; S = \text{HPTN } 084)$) as a sensitivity parameter. In this way, if we assume no cross-over in the hypothetical placebo-controlled trial (i.e., $CC_P(\mathbf{X}; S = \text{HPTN } 084) = 0$), then the intention-to-treat effect of daily oral TDF/FTC against placebo in the target population was estimated to be −3.9 HIV infections per 100 person-years (95% CI: −9.7 to −0.7). In a sensitivity analysis, we further allowed some minor degree of cross-over by setting $CC_P(\mathbf{X}; S = \text{HPTN } 084) = 5\%$ and 10%, and the ITT effect of TDF/FTC was estimated to be −3.5 per 100 person-years (95% CI: −8.7 to −0.6) and −3.1 per 100 person-years (95% CI: −7.8 to −0.5), respectively. We then conducted inference using the EIF-based estimator proposed in Section 4.2. The ITT effect of TDF/FTC was estimated to be −3.1 per 100 person-years (95% CI: −5.3, −0.9) assuming no cross-over. If we set $CC_P(\mathbf{X}; S = \text{HPTN084}) = 5\%$ and 10%, the ITT effect of TDF/FTC was estimated to be −2.5 per 100 person-years (95% CI: −4.5, −0.5) and −1.8 per 100 person-years (95% CI: −3.6, 0.0), respectively. In this integrated analysis, we found that the EIF-based estimators were more efficient compared to the regression-based estimators.

Because the target population and the Partners PrEP population was not well-overlapped in baseline commodities including Gonorrhea and Chlamydia, we further considered an analysis restricted to participants testing negative for Gonorrhea, Chlamydia and Trichomonas at baseline. For this target population, the ITT effect of TDF/FTC against placebo was estimated to be −3.3 per 100 person-years (95% CI: −9.0 to −0.4) assuming no cross-over. According to the EIF-based estimator, the ITT effect of TDF/FTC was estimated to be −2.2 per 100 person-years (95% CI: −4.5 to 0.0) assuming no cross-over.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Finally, we considered relaxing Assumption 3 in the Partners PrEP study and only partially identifying the conditional average treatment effect of daily TDF/FTC versus placebo. To this end, we considered the following strategy. Within each stratum defined by observed covariates **X**, the conditional average treatment effect can be decomposed into a weighted sum of the treatment effect among the subgroup of compliers and that among the non-compliers, weighted by their relatively proportions in the stratum. It then suffices to determine the average treatment effect among the compliers, non-compliers, and their proportions, all within the strata of observed covariates. Conditional on the observed covariates, the complier average effect is identified by the ratio estimator (4). On the other hand, the treatment effect of daily oral TDF/FTC among non-compliers has the following natural bounds: the maximum treatment effect was to reduce all HIV incidence in the placebo arm of the Partners PrEP study and the minimum effect was 0. In this way, we estimated bounds for the conditional average treatment effect and used these bounds to form the final ITT effect estimates against placebo in the target population. Assuming no cross-over, the interval estimates of the ITT effect were [−3.6, −2.5] per 100 person-years (95% CI: −8.0 to −0.4).

6.4 Estimating the absolute efficacy of CAB-LA in the target population

Our analysis also immediately implies that the HIV incidence was 6.5 (95% CI: 3.1 to 12.4) per 100 person-years in the counterfactual placebo arm (primary analysis under the point identification assumptions and based on the regression-based estimator) in the target population. This estimate became 5.5 (95% CI: 1.5 to 11.0]) per 100 person-years if we further restrict the target population to those who tested negative for Gonorrhea, Chlamydia and Trichomonas at baseline. These estimates of placebo arm HIV incidence agreed reasonably well with those reported in the FEM-PrEP study (5.0 per 100 person-years) and the VOICE study (4.6 per 100 person-years). On the other hand, under a naïve adoption of the constancy assumption, one would conclude an HIV incidence of 3.7 per 100 person-year in the target population, which appeared to largely underestimating the HIV incidence among young women in sub-Saharan Africa. Our result also implies an absolute efficacy of CAB-LA as large as −6.1 per 100 person-years (95% CI: −11.9 to −2.6) in the target population and −5.3 per 100 person-years (95% CI: −10.5 to −1.3) if the target population was further restricted to those who tested negative for Gonorrhea, Chlamydia and Trichomonas at baseline. Put together the estimates for HIV incidence in the placebo arm and the estimates of the efficacy of CAB-LA, *we estimated that CAB-LA eliminated about 95% of HIV acquisitions in the target population*.

7 Discussion

In this article, we systematically study the problem of generalizing the intention-to-treat effect of an active control versus placebo from historical placebo-controlled trials to an active-controlled trial. Our key insight is that generalization critically depends on the post-randomization event like adherence to the prescribed treatment in clinical trials. Our framework helps translate what FDA refers to as non-statistically-based uncertainties (Food and Drug Administration, 2016, Page 20) into concrete causal identification assumptions, highlights multiple sources of heterogeneity, including heterogeneity in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

participants composition, compliance, and treatment effect, and emphasizes the role of a post-randomization event when generalizing and transporting causal conclusions.

Our work adds to existing HIV prevention literature on inferring intent-to-treat effect of an active-control based on a "counterfactual placebo" incidence estimate, which may be constructed via leveraging data from a concurrent registrational cohort that receives access to available standard of care for HIV prevention (US National Library of Medicine, 2021), placebo arm data of historical trials in a similar population (Donnell et al., 2022), HIV recency testing data collected at screening (Gao et al., 2021) and adherence-efficacy relationship (Glidden et al., 2020, 2021).

The statistical problem of generalizing the ITT effect of an active control versus placebo from relevant historical trials has become more relevant as active-controlled trials have become increasingly prevalent. There are several ways to further this line of research. One important future direction is to generalize the framework to more complicated settings where post-randomization events like adherence to the intervention are time-varying, and the endpoint of interest is a time-to-event endpoint. Second, compliance or adherence to the intervention in an instrumental variable framework is a particular instance of a post-randomization event. It is also of interest to further extend the framework to a more generic, post-randomization event and allow a direct effect from the treatment to the endpoint of interest. Lastly, in many practical circumstances, researchers may not have the luxury to work with the patient-level data across multiple phase 3 clinical trials. Study-level adherence from historical trials has been used in a meta-regression analysis to infer oral PrEP effectiveness (Hanscom et al., 2019). Other meta-analysis-based approaches are also available; see, e.g., related discussion in Section 1.3. It is of interest to link the patient-level analysis proposed in this article to the meta-analysis-based framework and articulate what identification and modeling assumptions are needed to facilitate using only summary data from relevant historical trials.

Two statistical challenges are particularly relevant in generalizing efficacy estimates from historical data. First, researchers need to always pay close attention to the overlapping covariate space between the planned active-controlled trial and historical trials and, in our opinion, should always focus on the well-overlapped covariate space to avoid over-extrapolation with limited data. Traditional methods like multivariate matched sampling (Rubin, 1979) can be generalized to the context of across-trials comparisons; see, e.g., Zhang (2023). Examining the scalar summary statistic like Stuart et al.'s (2011) "probability of participation" is also useful. If the target trial enrolls a heterogeneous population of participants, then it is conceivable that multiple historical trials targeting different different constituent parts of the target population may be needed. Second, in some cases, it is conceivable that trials may not maintain a similar list of important covariates or may collect different versions of the same covariates. This is less of a concern if the studies were conducted via the same clinical trials network (e.g., the HIV Prevention Trials Network and the HIV Vaccine Trials Network) but could lead to many practical challenges and prevent researchers from pursuing covariate adjustment in other cases. Classical measurement error methods or methods that leverage proxy variables could be useful.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We appreciate all the constructive feedback from the editor, associate editor, and two anonymous reviewers. We are grateful to the study participants, study staff and investigators on HPTN 084 and Partners PrEP who provided the data for this analysis. We acknowledge the funders and sponsors of the trials. We are grateful to the HPTN Manuscript Review Committee for helpful feedback. This work was supported by the U.S. National Institutes of Health grants R01AI177078 and UM1AI068617 (Fei Gao) and by the VIDD Faculty Initiative Award at the Fred Hutchinson Cancer Center (Fei Gao and Bo Zhang). Oliver Dukes received support from the Research Foundation Flanders (1222522N).

References

Angrist JD, Imbens GW, and Rubin DB (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434):444–455.

Baeten JM, Donnell D, Ndase P, Mugo NR, Campbell JD, Wangisi J, Tappero JW, Bukusi EA, Cohen CR, Katabira E, et al. (2012). Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. New England Journal of Medicine, 367(5):399–410. [PubMed: 22784037]

Bickel PJ, Klaassen CA, Bickel PJ, Ritov Y, Klaassen J, Wellner JA, and Ritov Y (1993). Efficient and adaptive estimation for semiparametric models, volume 4. Springer.

Cheng G and Huang JZ (2010). Bootstrap consistency for general semiparametric m-estimation. The Annals of Statistics, 38(5):2884–2915.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, and Newey W (2017). Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5):261–65.

Cohen MS and Baden LR (2012). Preexposure prophylaxis for HIV—where do we go from here? New England Journal of Medicine, 367(5):459–461. [PubMed: 22784041]

Cole SR and Stuart EA (2010). Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. American journal of epidemiology, 172(1):107–115. [PubMed: 20547574]

Dahabreh IJ, Robertson SE, and Hernán MA (2022). Generalizing and transporting inferences about the effects of treatment assignment subject to non-adherence. arXiv preprint arXiv:2211.04876.

Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, and Hernán MA (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics, 75(2):685–694. [PubMed: 30488513]

Degtiar I and Rose S (2021). A review of generalizability and transportability. arXiv preprint arXiv:2102.11904.

Delany-Moretlwe S, Hughes JP, Bock P, Ouma SG, Hunidzarira P, Kalonji D, Kayange N, Makhema J, Mandima P, Mathew C, et al. (2022). Cabotegravir for the prevention of HIV-1 in women: results from HPTN 084, a phase 3, randomised clinical trial. The Lancet, 399(10337):1779–1789.

Donnell D, Gao F, Hughes J, and Hanscom B (2022). Counterfactual estimation of CAB-LA efficacy against placebo using external trials. volume 86, Virtual.

Ellenberg SS and Temple R (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 2: practical issues and specific cases. Annals of Internal Medicine, 133(6):464–470. [PubMed: 10975965]

Fauci AS (2017). An hiv vaccine is essential for ending the hiv/aids pandemic. Jama, 318(16):1535–1536. [PubMed: 29052689]

Fleming TR, Odem-Davis K, Rothmann MD, and Li Shen Y (2011). Some essential considerations in the design and conduct of non-inferiority trials. Clinical Trials, 8(4):432–439. [PubMed: 21835862]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Food and Drug Administration (2016). Non-inferiority clinical trials to establish effectiveness: Guidance for industry.

Gao F, Glidden DV, Hughes JP, and Donnell DJ (2021). Sample size calculation for active-arm trial with counterfactual incidence based on recency assay. Statistical Communications in Infectious Diseases, 13(1).

Glidden DV, Das M, Dunn DT, Ebrahimi R, Zhao Y, Stirrup OT, Baeten JM, and Anderson PL (2021). Using the adherence-efficacy relationship of emtricitabine and tenofovir disoproxil fumarate to calculate background hiv incidence: a secondary analysis of a randomized, controlled trial. Journal of the International AIDS Society, 24(5):e25744. [PubMed: 34021709]

Glidden DV, Stirrup OT, and Dunn DT (2020). A bayesian averted infection framework for prep trials with low numbers of hiv infections: application to the results of the discover trial. The Lancet HIV, 7(11):e791–e796. [PubMed: 33128906]

Grant RM, Anderson PL, McMahan V, Liu A, Amico KR, Mehrotra M, Hosek S, Mosquera C, Casapia M, Montoya O, et al. (2014). Uptake of pre-exposure prophylaxis, sexual practices, and hiv incidence in men and transgender women who have sex with men: a cohort study. The Lancet infectious diseases, 14(9):820–829. [PubMed: 25065857]

Hahn J (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica, pages 315–331.

Hanscom B, Hughes JP, Williamson BD, and Donnell D (2019). Adaptive non-inferiority margins under observable non-constancy. Statistical methods in medical research, 28(10–11):3318–3332. [PubMed: 30293490]

Hastie TJ (2017). Generalized additive models. In Statistical models in S, pages 249–307. Routledge.

James Hung H, Wang S-J, Tsong Y, Lawrence J, and O'Neil RT (2003). Some fundamental issues with non-inferiority testing in active controlled trials. Statistics in Medicine, 22(2):213–225. [PubMed: 12520558]

Joffe MM and Greene T (2009). Related causal frameworks for surrogate outcomes. Biometrics, 65(2): 530–538. [PubMed: 18759836]

Marrazzo JM, Ramjee G, Richardson BA, Gomez K, Mgodi N, Nair G, Palanee T, Nakabiito C, Van Der Straten A, Noguchi L, et al. (2015). Tenofovir-based preexposure prophylaxis for HIV infection among African women. New England Journal of Medicine, 372(6):509–518. [PubMed: 25651245]

Miner MD, Corey L, and Montefiori D (2021). Broadly neutralizing monoclonal antibodies for hiv prevention. Journal of the International AIDS Society, 24:e25829. [PubMed: 34806308]

Murnane PM, Brown ER, Donnell D, Coley RY, Mugo N, Mujugira A, Celum C, Baeten JM, Team PPS, Mujugira A, et al. (2015). Estimating efficacy in a randomized trial with product nonadherence: application of multiple methods to a trial of preexposure prophylaxis for HIV prevention. American Journal of Epidemiology, 182(10):848–856. [PubMed: 26487343]

Neilan AM, Landovitz RJ, Le MH, Grinsztejn B, Freedberg KA, McCauley M, Wattananimitgul N, Cohen MS, Ciaranello AL, Clement ME, et al. (2022). Cost-effectiveness of long-acting injectable hiv preexposure prophylaxis in the united states: a cost-effectiveness analysis. Annals of internal medicine, 175(4):479–489. [PubMed: 35099992]

Pearl J (2011). Transportability across studies: A formal approach.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55.

Rothmann M, Li N, Chen G, Chi GY, Temple R, and Tsou H-H (2003). Design and analysis of non-inferiority mortality trials in oncology. Statistics in Medicine, 22(2):239–264. [PubMed: 12520560]

Rubin D (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. Journal of the American Statistical Association, 75:591–593.

Rubin DB (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association, 74(366a):318–328.

Rudolph KE and van der Laan MJ (2017). Robust estimation of encouragement design intervention effects transported across sites. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(5):1509–1525. [PubMed: 29375249]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Stefanski LA and Boos DD (2002). The calculus of M-estimation. The American Statistician, 56(1):29–38.

Stuart EA, Cole SR, Bradshaw CP, and Leaf PJ (2011). The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society. Series A, (Statistics in Society), 174(2):369–386.

Swanson SA, Hernán MA, Miller M, Robins JM, and Richardson TS (2018). Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. Journal of the American Statistical Association, 113(522):933–947. [PubMed: 31537952]

US National Library of Medicine (2021). A Combination Efficacy Study in Africa of Two DNA-MVA-Env Protein or DNA-Env Protein HIV-1 Vaccine Regimens With PrEP (PrEPVacc).

Van Damme L, Corneli A, Ahmed K, Agot K, Lombaard J, Kapiga S, Malahleha M, Owino F, Manongi R, Onyango J, et al. (2012). Preexposure prophylaxis for HIV infection among African women. New England Journal of Medicine, 367(5):411–422. [PubMed: 22784040]

van der Laan MJ and Rose S (2011). Targeted learning: causal inference for observational and experimental data, volume 10. Springer.

Van der Vaart AW (2000). Asymptotic statistics, volume 3. Cambridge university press.

Wang L and Tchetgen Tchetgen E (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(3):531–550. [PubMed: 30034269]

World Health Organization (2022). Guidelines on long-acting injectable cabotegravir for HIV prevention. World Health Organization.

Zhang B (2023). Efficient algorithms for building representative matched pairs with enhanced generalizability. Biometrics (in press).

Zhang Z (2009). Covariate-adjusted putative placebo analysis in active-controlled clinical trials. Statistics in Biopharmaceutical Research, 1(3):279–290.

Zhang Z, Nie L, Soon G, and Zhang B (2014). Sensitivity analysis in non-inferiority trials with residual inconstancy after covariate adjustment. Journal of the Royal Statistical Society: Series C (Applied Statistics), 63(4):515–538.

Zhou J, Hodges JS, Suri MFK, and Chu H (2019). A bayesian hierarchical model estimating cace in meta-analysis of randomized clinical trials with noncompliance. Biometrics, 75(3):978–987. [PubMed: 30690716]

Zhou T, Zhou J, Hodges JS, Lin L, Chen Y, Cole SR, and Chu H (2022). Estimating the complier average causal effect in a meta-analysis of randomized clinical trials with binary outcomes accounting for noncompliance: A generalized linear latent and mixed model approach. American journal of epidemiology, 191(1):220–229. [PubMed: 34564720]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

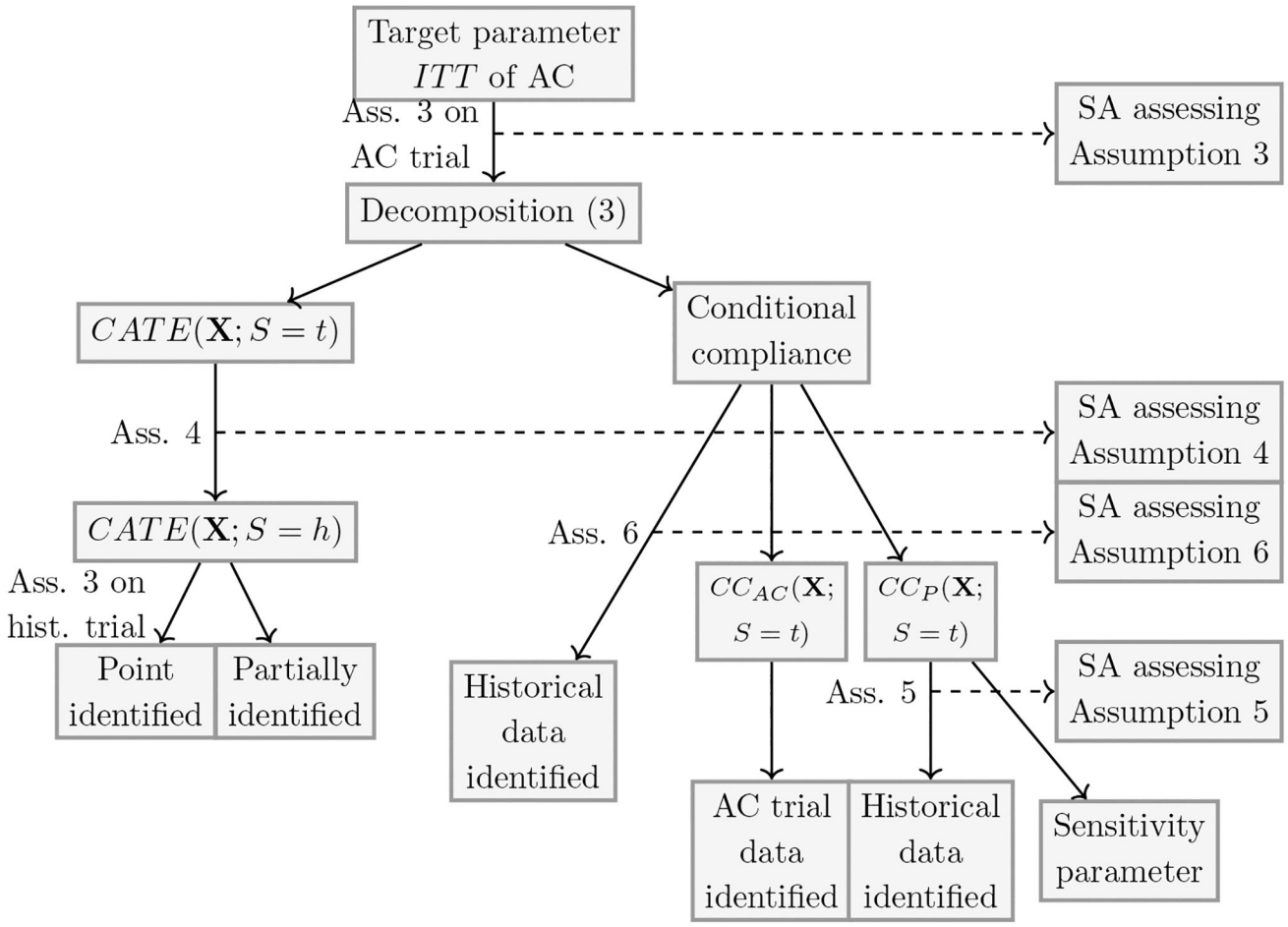

Figure 1:
A schematic flow chart summarizing different identification assumptions, quantities involved in the estimation, mode of identification, and associated sensitivity analyses examining core assumptions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

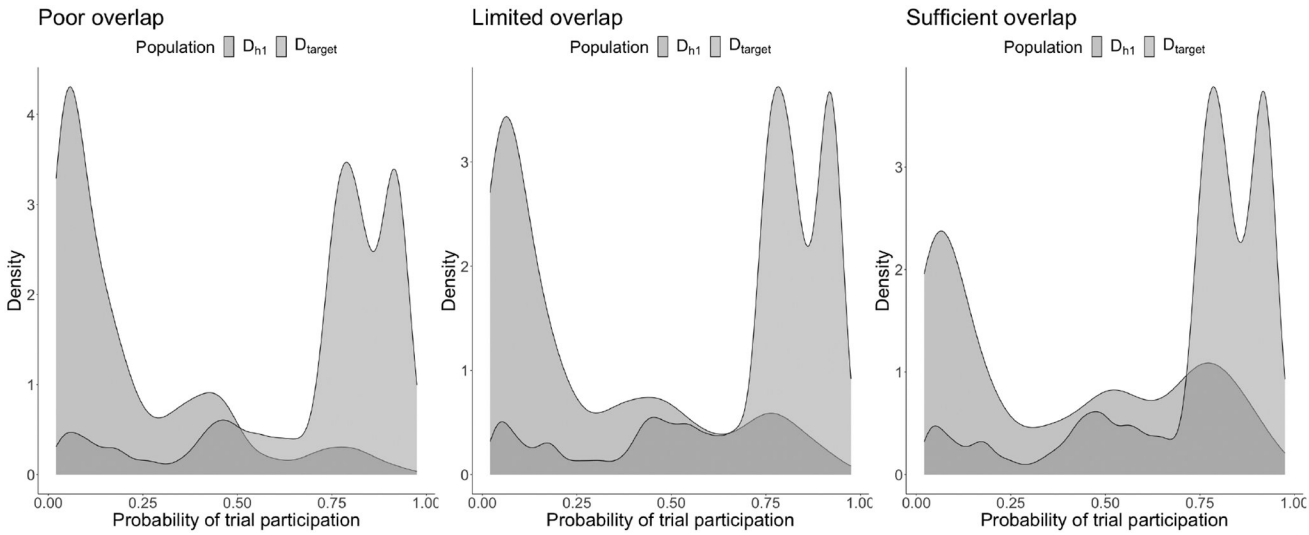

Figure 2:
The probability of trial participation based on a participant's baseline characteristics in poor, limited, and sufficient overlap scenarios.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

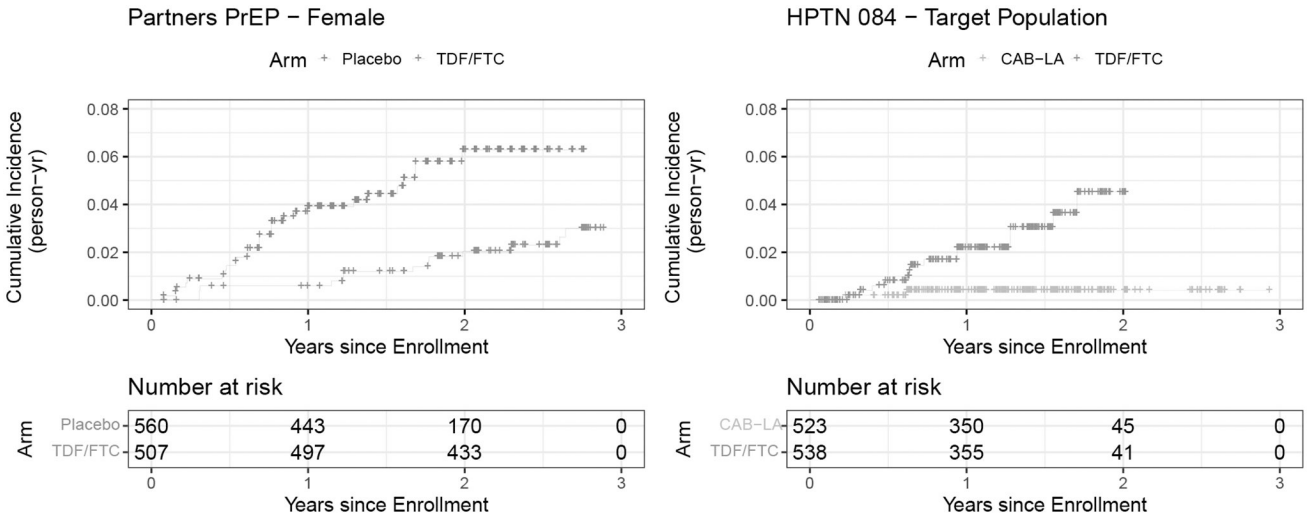

Figure 3:

Left panel: Kaplan-Meier estimates of incident HIV acquisition in the Partners PrEP study. Right panel: Kaplan-Meier estimates of incident HIV acquisition among HPTN 084 participants whose partners either living with HIV or having an unknown HIV status (target population).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1:

Simulation results of 6 estimators corresponding to Scenario X1 and Scenario Y1. The percentage of bias and coverage of 95% confidence intervals are reported. Confidence intervals of $\widehat{ITT}_{\text{EIF, gam}}$ were estimated based on asymptotic normality and the efficient influence function. Confidence intervals of $\widehat{ITT}_{\text{hypo}}$ were based on two-sample tests. Confidence intervals of the other estimators were obtained via bootstrap. Out of 1,000 simulations, $\widehat{ITT}_{\text{EIF, gam}}$ fell outside [−1, 1] twice in the limited overlap setting and once in the sufficient overlap setting when $n = 1{,}000$.

	$\widehat{ITT}_{\text{hypo}}$		$\widehat{ITT}_{\text{const, 1}}$		$\widehat{ITT}_{\text{const, 2}}$		$\widehat{ITT}_{\text{reg, par}}$		$\widehat{ITT}_{\text{EIF, par}}$		$\widehat{ITT}_{\text{EIF, gam}}$	
Sample size	**% Bias**	**95% CI Coverage**	**% Bias**	**95% CI Coverage**	**% Bias**	**95% CI Coverage**	**% Bias**	**95% CI Coverage**	**% Bias**	**95% CI Coverage**	**% Bias**	**95% CI Coverage**
Poor Overlap												
1000	0.0	96.0%	−16.5	93.2%	73.2	27.0%	2.5	96.0%	2.5	96.2%	2.7	84.0%
2000	−0.1	97.4%	−17.8	87.8%	72.1	5.0%	0.8	95.6%	−0.7	92.8%	−0.5	85.8%
5000	−0.1	96.4%	−17.0	81.0%	70.4	0.0%	2.2	96.0%	2.3	94.9%	2.3	88.1%
Limited Overlap												
1000	−0.6	96.2%	−16.1	91.2%	72.4	19.2%	3.2	93.4%	1.6	94.4%	2.8	89.6%
2000	0.4	97.2%	−17.5	85.6%	70.6	3.8%	1.3	95.6%	0.8	94.6%	0.8	88.6%
5000	0.3	95.3%	−17.0	73.2%	72.4	0.0%	2.0	94.7%	2.3	95.1%	2.3	90.4%
Sufficient Overlap												
1000	0.1	96.2%	−15.9	88.2%	72.2	5.6%	3.9	94.6%	3.4	94.6%	4.2	92.8%
2000	−0.7	96.4%	−16.6	83.0%	73.2	0.0%	2.7	94.0%	2.7	93.8%	3.0	90.8%
5000	0.1	95.8%	−17.1	62.2%	71.9	0.0%	1.8	94.7%	2.0	95.2%	2.0	93.6%

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2:

Baseline characteristics of all participants in HPTN 084, HPTN 084 participants whose partners lived with HIV, and female participants in the Partners PrEP study. Mean (SD) are reported for continuous variables. Counts (%) are reported for categorical variables.

	HPTN 084 All participants	HPTN 084 Participants whose partner lived with HIV or had an unknown status	Partners PrEP Female participants whose partners lived with HIV
	(N=3224)	(N=1139)	(N=1184)
Study arm			
CAB-LA	1614 (50.1%)	559 (49.1%)	0 (0%)
TDF/FTC	1610 (49.9%)	580 (50.9%)	565 (47.7%)
Placebo	0 (0%)	0 (0%)	619 (52.3%)
Age	26.0 (5.78)	26.3 (6.03)	33.5 (7.55)
Gonorrhea			
Neg	2977 (92.3%)	1059 (93.0%)	1068 (90.2%)
Pos	210 (6.5%)	68 (6.0%)	14 (1.2%)
Missing	37 (1.1%)	12 (1.1%)	102 (8.6%)
Chlamydia			
Neg	2583 (80.1%)	941 (82.6%)	1068 (90.2%)
Pos	604 (18.7%)	186 (16.3%)	13 (1.1%)
Missing	37 (1.1%)	12 (1.1%)	103 (8.7%)
Trichomonas			
Neg	2859 (88.7%)	1021 (89.6%)	1057 (89.3%)
Pos	270 (8.4%)	89 (7.8%)	80 (6.8%)
Missing	95 (2.9%)	29 (2.5%)	47 (4.0%)
Syphilis			
Neg	3116 (96.7%)	1107 (97.2%)	1101 (93.0%)
Pos	103 (3.2%)	30 (2.6%)	69 (5.8%)
Missing	5 (0.2%)	2 (0.2%)	14 (1.2%)
Employment			
Employed	878 (27.2%)	276 (24.2%)	807 (68.2%)
Not employed	2346 (72.8%)	863 (75.8%)	377 (31.8%)
Education			
Complete secondary school	1528 (47.4%)	527 (46.3%)	75 (6.3%)
Not complete secondary school	1346 (41.7%)	506 (44.4%)	271 (22.9%)
Not complete primary school	350 (10.9%)	106 (9.3%)	838 (70.8%)