

Addressing selection bias in cluster randomized experiments via weighting

Georgia Papadogeorgou¹, Bo Liu², Fan Li³, Fan Li²

¹Department of Statistics, University of Florida, USA

²Department of Statistical Science, Duke University, USA

³ Department of Biostatistics, Yale University, USA

Abstract

In cluster randomized experiments, individuals are often recruited after the cluster treatment assignment, and data are typically only available for the recruited sample. Post-randomization recruitment can lead to selection bias, inducing systematic differences between the overall and the recruited populations, and between the recruited intervention and control arms. In this setting, we define causal estimands for the overall and the recruited populations. We prove, under the assumption of ignorable recruitment, that the average treatment effect on the recruited population can be consistently estimated from the recruited sample using inverse probability weighting. Generally we cannot identify the average treatment effect on the overall population. Nonetheless, we show, via a principal stratification formulation, that one can use weighting of the recruited sample to identify treatment effects on two meaningful subpopulations of the overall population: individuals who would be recruited into the study regardless of the assignment, and individuals who would be recruited into the study under treatment but not under control. We develop an estimation strategy and a sensitivity analysis approach for checking the ignorable recruitment assumption. The proposed methods are applied to the ARTEMIS cluster randomized trial, where removing co-payment barriers increases the persistence of P2Y₁₂ inhibitor among the always-recruited population.

keywords: Causal inference; Cluster randomized trial; Principal stratification; Sensitivity analysis; Selection bias; Working propensity score

1. Introduction

Randomized experiments are the gold standard for evaluating causal effects. Randomization guarantees that the treatment arms are comparable in both measured and unmeasured baseline covariates. However, when the inclusion of an individual unit, which we generically refer to as *recruitment* or *enrollment* herein, into an experiment is determined after the treatment assignment, individuals can self-select whether to participate in the experiment, inducing post-randomization selection bias. This problem is prevalent in cluster randomized experiments, where individuals are typically

recruited after cluster randomization and are not blinded to the assignment. Patients are often more willing to participate in the study if they are under a specific treatment arm, especially when that treatment has a perceived benefit. Consequently, the recruited treatment and control arms can be systematically different, breaking the initial randomization.

We consider post-randomization selection as a special case of the general setting of an intermediate variable lying temporally between an exposure and an outcome [Rosenbaum, 1984, Frangakis and Rubin, 2002]. The selection bias setting due to recruitment is unique in that data on the unrecruited individuals are entirely missing, whereas in many other settings we have at least partial data on all individuals. Post-randomization selection bias is particularly prevalent in cluster randomized experiments where individual recruitment occurs after cluster-level treatment assignment, and blinding is frequently impractical. Also known as recruitment or identification bias, this type of selection bias was recognized by Puffer et al. [2003] in a review of cluster randomized experiments published from 1997 to 2002. Brierley et al. [2012] conducted a review of 24 published cluster randomized experiments in four leading medical journals and found that 40% identified individuals after randomization and could be subject to selection bias. Similar observations have been repeatedly confirmed in systematic reviews of cluster randomized experiments, for example, by Bolzern et al. [2018] and Easter et al. [2021]. Due to the relevance of selection bias in the cluster randomized setting, we address a fundamental question—that is, *whether one can estimate meaningful causal estimands on the overall population based on the recruited sample alone*. Our setting is different from the challenges arising in randomized experiments due to selective drop-out when the treatment assignment is not blinded or it can be inferred based on potential side effects [Little and Yau, 1996], where data are typically available on the whole population.

Post-randomization selection has important implications in the analysis of cluster randomized experiments [Li et al., 2022a], and if ignored, often leads to ambiguous causal estimands and estimation bias. First, the recruited sample is generally not a simple random sample of the overall population; therefore, it is important to differentiate between the causal estimands defined on the recruited population and the overall population. Although previous simulations have recommended regression adjustment to address post-randomization selection bias in cluster randomized experiments [Leyrat et al., 2013, 2014], they have restricted to an overly strong homogeneous treatment effect assumption that masks the importance of target population and estimands. Second, in the presence of post-randomization selection, one generally cannot estimate well-defined causal estimands on the recruited population without additional assumptions [Wang et al., 2024, Schochet,

2024]. For example, Wang et al. [2024] considered a scenario where the recruited cluster size depends on the treatment, cluster-level covariates as well as the underlying cluster population size. However, they did not address a likely scenario where the recruitment process further depends on individual-level characteristics. Finally, current methodology suggests that it is necessary to obtain additional data on at least a portion of the non-recruited individuals to validly estimate causal effects on the overall population [Li et al., 2022b]. However, such additional data are usually not directly available in a given experiment.

In this article, we clarify the causal estimands, contribute novel identification results, and develop weighting methods that leverage individual-level covariates to address post-randomization selection bias in cluster randomized experiments, without assuming homogeneous treatment effects nor additional information among the non-recruited. We show that, under an ignorability and a monotonicity assumption on the recruitment mechanism, one can nonparametrically identify, using only the recruited sample, the average treatment effect on (i) the recruited population, and (ii) two interpretable subpopulations of the overall population. The central tool is weighting by the *working propensity score*, which is the conditional probability of being in the treatment arm among the recruited. Recognizing that the ignorable recruitment assumption may not always hold, we then develop a new sensitivity analysis method to assess the impact under violations of this assumption.

The proposed methods are motivated by the ARTEMIS cluster randomized trial [the Affordability and Real-World Antiplatelet Treatment Effectiveness After Myocardial Infarction Study; Wang et al., 2019]. The ARTEMIS trial aims at evaluating whether removing co-payment barriers increases the persistence of P2Y₁₂ inhibitor—a common drug for patients who had experienced myocardial infarction—and lowers risk of major adverse cardiovascular events. The intervention is randomized at the hospital level, and patients were recruited in the study with knowledge of the hospital-level treatment assignment. Only the intervention hospitals offered vouchers for P2Y₁₂ to the enrolled patients with no out-of-pocket cost. Because the intervention is an obvious incentive to the patients, enrolling patients in the control hospitals was more challenging than the intervention hospitals, leading to potential selection bias due to recruitment. Using our methods, we identify a positive causal effect of removing co-payment barriers on persistence of P2Y₁₂ inhibitor in the patient subpopulation that would be recruited under either treatment arm. Furthermore, in ARTEMIS, since individuals in the treated clusters were readily recruited, the patient population in the treated clusters is likely representative of the overall population. We report an increase in medication persistence on this treated population, which is suggestive of a potential positive effect

of removing co-payment barriers on the overall population.

2. Notation and Causal Estimands

Consider a cluster randomized experiment which consists of J clusters drawn from a super-population of clusters, among which m clusters are randomized to the intervention arm, denoted by $Z_j^c = 1$, and the remaining clusters to the control arm, denoted by $Z_j^c = 0$. The superscript ‘c’ indicates a cluster-level variable. Even though treatment randomization is considered the gold-standard for drawing causal inferences, data from cluster randomized experiments can suffer from selection bias stemming from individual decisions on cluster membership and enrollment after the cluster-level treatment assignment [Schochet, 2024].

Whether selection bias due to cluster membership takes place depends on whether the treatment assignment affects the individual choice of cluster. For example, in health care of rare diseases, patients might choose a hospital that is specialized in treating their conditions or offers a new experimental treatment that is perceived as beneficial. Such decision making process on individual cluster membership would render treated and control clusters imbalanced in their composition and would likely lead to a larger load of patients in treated hospitals. We do not consider this scenario and refer to Schochet [2024] for more details. Instead, we focus on the scenario where individuals do not self-select into a cluster based on the cluster-level treatment assignment. For example, individuals in a classroom or village are usually fixed at the time of randomization and would remain in that cluster throughout the study. Another example is health care of common diseases such as heart disease or diabetes; patients usually choose a hospital based on distance or insurance and would not change hospitals. This is the case in our study (Section 7). We formalize this scenario in Assumption 1.

Assumption 1. (Constant cluster membership) *Let $\mathcal{J}, \mathcal{J}'$ denote two sets of clusters that could comprise the cluster randomized experiment, and $\mathbf{z} \in \{0, 1\}^{|\mathcal{J}|}, \mathbf{z}' \in \{0, 1\}^{|\mathcal{J}'|}$ two hypothetical cluster treatment assignments in each composition, respectively. Let $Q(\mathcal{J}, \mathbf{z})$ denote the cluster that a randomly chosen individual would belong to under cluster composition \mathcal{J} and treatment assignment \mathbf{z} . We assume that $Q(\mathcal{J}, \mathbf{z}) = Q(\mathcal{J}', \mathbf{z}')$.*

Under Assumption 1, if we conceive that individuals form a super-population and they arrive at the super-population of clusters under some random process, then this process is constant with respect to the study, and the individuals in the cluster randomized experiment can be considered

fixed. We refer to all individuals, recruited or not, across the clusters in the experiment as the *overall population*. Assumption 1 also allows us to use a double-index notation: ‘ ij ’ denotes individual i in cluster j with the total number of individuals N_j . Under cluster randomization, we also have $Z_{ij} = Z_j^c$. We use $N = \sum_j N_j$ to denote the number of individuals in the overall population.

The second source of selection bias in cluster randomized experiments arises from selective enrollment or recruitment into the study of individuals in the overall population that are eligible for the experiment. Selective recruitment is common in cluster randomized experiments in the healthcare delivery context across disease conditions, as suggested by a series of reviews [Puffer et al., 2003, Brierley et al., 2012, Bolzern et al., 2018], and has been found to be particularly common in cluster randomized studies evaluating individual-level interventions [Easter et al., 2021]. Mechanistically, after clusters are randomized, individuals from treated clusters may be more likely to give consent to join the study when the intervention is perceived to be beneficial, or recruiters from clusters may be more motivated to enroll individuals based on their foreknowledge of cluster assignment. In either case, the subpopulation recruited into the study (denoted by $R_{ij} = 1$) is referred to as the *recruited population*, which is different from the overall population as long as some individuals are unrecruited (denoted by $R_{ij} = 0$). We denote the number of recruited individuals in cluster j by $n_j \leq N_j$, and the total number of recruited individuals in the study by $n = \sum_j n_j$.

Let V_{ij} and V_j^c denote the covariates for individual i in cluster j , and for cluster j , respectively, and write $X_{ij} = (V_{ij}, V_j^c)$ as the collection of covariates for individual i . Under the cluster-level Stable Unit Treatment Value Assumption, there are two individual-level potential outcomes $Y_{ij}(0)$ and $Y_{ij}(1)$ as well as two potential recruitment statuses, $R_{ij}(0)$ and $R_{ij}(1)$. We only observe the covariates and the potential outcomes corresponding to the observed assignment of the recruited individuals, denoted as $Y_{ij} = Y_{ij}(Z_{ij})$ for those with $R_{ij} = R_{ij}(Z_{ij}) = 1$. In what follows, we always use the notation ‘ \sum_{ij} ’ to denote a sum over the *recruited individuals* across all clusters.

We define the *average treatment effect on the overall population* as

$$\tau^O = E \{Y_{ij}(1) - Y_{ij}(0)\}.$$

We also define the *average treatment effect on the recruited population* and its counterpart specific

to treatment status $z \in \{0, 1\}$ as

$$\tau^R = E\{Y_{ij}(1) - Y_{ij}(0) \mid R_{ij} = 1\}; \quad \tau_z^R = E\{Y_{ij}(1) - Y_{ij}(0) \mid Z_{ij} = z, R_{ij} = 1\},$$

respectively. The expectation is over the cluster super-population, from which the cluster information $\{V_j^c, N_j, \{V_{ij}, R_{ij}(z), Y_{ij}(z)\}_{i=1}^{N_j}\}$ are independent and identically distributed draws. The estimand on the overall population τ^O and its counterparts on certain subpopulations of the overall population are usually the intended target estimands. However, researchers often resort to estimate τ^R because usually only data on the recruited sample are available. When the recruited sample is a simple random sample of the overall population or the treatment effect is homogeneous across all individuals, τ^O is equal to τ^R , but not so otherwise.

We now introduce causal estimands on meaningful subpopulations of the overall population via a principal stratification formulation [Frangakis and Rubin, 2002]. We classify the overall population by the joint individual potential recruitment statuses under both assignments: $S_{ij} = (R_{ij}(0), R_{ij}(1))$; S_{ij} is called a principal stratum. Due to the fundamental problem of causal inference, the individual principal stratum membership is not observable. We name the four principal strata as: $S_{ij} = (1, 1) \equiv a$, *always-recruited*, individuals who would be recruited regardless of the assignment; $S_{ij} = (0, 0) \equiv n$, *never-recruited*, individuals who would not be recruited regardless of the assignment; $S_{ij} = (0, 1) \equiv c$, *incentivized-recruited*, individuals who would be recruited under intervention but not under control; and $S_{ij} = (1, 0) \equiv d$, *disincentivized-recruited*, individuals who would be recruited under control but not under intervention. In the noncompliance literature [Angrist et al., 1996], the latter two groups are often referred to as compliers and defiers, respectively. By construction, principal stratum membership does not change by treatment assignment; therefore, we can define causal effects within each principal stratum, for example

$$\tau_s^O = E\{Y_{ij}(1) - Y_{ij}(0) \mid S_{ij} = s\}, \tag{1}$$

for $s \in \{a, n, c, d\}$. If π_s represents the proportion of principal stratum s in the overall population, then $\tau^O = \sum_s \pi_s \tau_s^O$. We also define causal effects on unions of principal strata, $\tau_{a,c}^O = E\{Y_{ij}(1) - Y_{ij}(0) \mid S_{ij} \in \{a, c\}\}$, which is equal to the weighted average of τ_a^O and τ_c^O with weights $\pi_a/(\pi_a + \pi_c)$ and $\pi_c/(\pi_a + \pi_c)$, respectively.

The relevance of an estimand for policy-making critically depends on the scientific context. In general, estimands on the recruited population can be hard to interpret since this population

corresponds to an imbalanced mixture of always-, incentivized-, and disincentivized-recruited individuals [Li et al., 2022a]. However, our study in Section 7 represents a situation where estimands on a subpopulation of the recruited population can be informative of general policy. In contrast, the causal effects defined over unions of principal strata weigh the strata populations according to their prevalence in the overall population, and therefore provide a more interpretable representation of the population. That said, since principal stratum membership is always unobserved, it is not immediately clear how to use strata-specific estimands for policy-making. In Section 7 we illustrate how one may design policy to target various principal strata by investigating their baseline covariate distributions.

3. Nonparametric Identification of Causal Estimands

3.1 Identification Assumptions

Li et al. [2022a] showed that a simple difference in means estimator using the recruited sample is generally biased for both τ^O and τ^R . In this section, we show how to nonparametrically identify the above causal estimands *based on the recruited sample alone*. We first need to formalize the random assignment of the clusters.

Assumption 2. (Cluster randomization). *Let $Z^c = (Z_1^c, Z_2^c, \dots, Z_J^c)$ denote the assignment of the clusters. Then, $\Pr(Z_j^c \mid Y_{ij}(0), Y_{ij}(1), V_{ij}, V_j^c \text{ for all } i, j) = \Pr(Z_j^c)$.*

Under Assumptions 1 and 2, we can prove that cluster randomization implies individual randomization, as stated in the next lemma (the proof is in Supplement B).

Lemma 1. *If Assumptions 1 and 2 hold, then the individual-level treatment assignment on the overall population is as-if randomized at the individual level, and the probability of treatment of an individual is equal to the cluster-level treatment probability, i.e. $\Pr(Z_{ij} = 1 \mid Y_{ij}(0), Y_{ij}(1), V_{ij}, V_j^c \text{ for all } i, j) = \Pr(Z_{ij} = 1) = \Pr(Z_j^c = 1)$.*

Assumption 1 characterizes the composition of the overall population. We now specify two related assumptions that characterize the composition of the recruited population.

Assumption 3A. (Non-differential recruitment) *The recruitment process is non-differential with respect to potential outcomes given covariates if there exists a function $\delta(x)$ such that*

$$\delta(x) = \frac{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 0)}, \quad (2)$$

for all x, y_0 , and y_1 .

Assumption 3B. (Ignorable recruitment) *Conditional on covariates and the treatment assignment, an individual’s recruitment status is independent of the potential outcomes:* $\Pr(R_{ij} = 1 \mid Y_{ij}(0), Y_{ij}(1), Z_{ij}, X_{ij}) = \Pr(R_{ij} = 1 \mid Z_{ij}, X_{ij})$.

Assumption 3A states that all factors driving the recruitment *differently* between treated and control individuals are measured, whereas Assumption 3B states that all factors driving the recruitment are measured, i.e. there is no unmeasured confounding with respect to the recruitment process. Assumption 3B implies Assumption 3A, but not vice versa. As we explain later, the function $\delta(x)$ is closely related to the proportion of individuals that belong to the different principal strata. Assumption 3A is equivalent to the subset ignorability in Gaebler et al. [2022]: $Z_{ij} \perp\!\!\!\perp \{Y_{ij}(0), Y_{ij}(1)\} \mid \{X_{ij}, R_{ij} = 1\}$. Since the treatment assignment temporally precedes the recruitment, it is arguably hard to justify subset ignorability in our setting because it imposes the independence of the treatment assignment conditional on the post-assignment recruitment. The ignorability assumption in Schochet [2024] carries the same issue in terms of the variables’ temporal ordering. Meanwhile, Assumption 3B is equivalent to conditional independence of the principal stratum membership and potential outcomes [referred to as principal ignorability by Ding and Lu, 2017]: $S_{ij} \perp\!\!\!\perp \{Y_{ij}(0), Y_{ij}(1)\} \mid X_{ij}$. (The relationship between the different assumptions is formally stated and proven in Supplement B.2.) For $\delta(x)$ to be well-defined, Assumption 3A is necessary. From here onwards, when we refer to $\delta(x)$, Assumption 3A is implicitly assumed to hold.

We maintain a standard positivity assumption that ensures that the probability of recruitment of any individual in the overall population is bounded away from 0.

Assumption 4 (*Recruitment positivity*). *There exists $\delta > 0$ such that $\Pr(R_{ij} = 1 \mid Z_{ij} = z, X_{ij} = x) > \delta$ for all (z, x) .*

3.2 Nonparametric Identification of Recruited and Overall Estimands

We first introduce the concept of a *working propensity score*, defined as the probability of being in the treatment group among the recruited: $e(x) = \Pr(Z_{ij} = 1 \mid X_{ij} = x, R_{ij} = 1)$. Importantly, the working propensity score does *not* reflect the true treatment assignment mechanism [Rosenbaum and Rubin, 1983], but is rather a one-dimensional summary of the measured covariates of the recruited individuals. Its true value depends on the conditional distribution of the recruitment

process through

$$e(x) = \frac{\Pr(R_{ij} = 1 \mid Z_{ij} = 1, X_{ij} = x)}{\Pr(R_{ij} = 1 \mid X_{ij} = x)} \Pr(Z_{ij} = 1),$$

which is unknown and cannot be estimated without data on the un-recruited. We discuss how to estimate the working propensity score in Section 4. If Assumption 4 holds, then the working propensity score is bounded away from 0 and 1 (see Supplement B.4).

With the recruited sample, τ^R is identifiable using inverse probability weighting, as

$$\tau^R = \mathbb{E} \left[\frac{Z_{ij}Y_{ij}}{e(X_{ij})} - \frac{(1 - Z_{ij})Y_{ij}}{1 - e(X_{ij})} \mid R_{ij} = 1 \right]. \quad (3)$$

Here, Assumption 3A suffices in place of the stronger Assumption 3B. The proof is in Supplement B.5. Note that τ^R is also identifiable using the standard outcome modeling strategy. The identification formula for τ_z^R is similar, with the contribution of each individual multiplied by $e(X_{ij})$ for τ_1^R and by $1 - e(X_{ij})$ for τ_0^R [Li et al., 2018].

We next show how to identify causal effects on two meaningful subpopulations of the overall population from the recruited sample alone. Notice that the observed cells of (R, Z) consists of mixtures of principal strata. Specifically, the recruited individuals from the intervention clusters $(Z_{ij} = 1, R_{ij} = 1)$ consist of always- and incentivized-recruited, whereas the recruited individuals from the control clusters $(Z_{ij} = 0, R_{ij} = 1)$ consist of always- and disincentivized-recruited. This observation connects two estimands for the overall population to estimands for the recruited population, summarized in Theorem 1.

Theorem 1. *If Assumptions 1, 2 and 3B hold, we have (a) $\tau_{a,d}^O = \tau_0^R$, and (b) $\tau_{a,c}^O = \tau_1^R$. If Assumption 4 also holds, the average treatment effect on the always- and disincentivized-recruited subpopulation of the overall population, $\tau_{a,d}^O$, is identifiable as*

$$\tau_{a,d}^O \equiv \mathbb{E}[Y_{ij}(1) - Y_{ij}(0) \mid S_{ij} \in \{a, d\}] = \mathbb{E} \left[\frac{Z_{ij}Y_{ij}\{1 - e(X_{ij})\}}{e(X_{ij})} - (1 - Z_{ij})Y_{ij} \mid R_{ij} = 1 \right], \quad (4)$$

and the average treatment effect among the always- and incentivized-recruited subpopulation of the overall population, $\tau_{a,c}^O$, is identifiable as

$$\tau_{a,c}^O \equiv \mathbb{E}[Y_{ij}(1) - Y_{ij}(0) \mid S_{ij} \in \{a, c\}] = \mathbb{E} \left[Z_{ij}Y_{ij} - \frac{(1 - Z_{ij})Y_{ij}e(X_{ij})}{1 - e(X_{ij})} \mid R_{ij} = 1 \right]. \quad (5)$$

Theorem 1 states that the causal effect on the union of always- and disincentivized-recruited in the overall population is equal to the causal effect on the recruited control arm, and that the causal effect on the union of the always- and incentivized-recruited in the overall population is equal to the effect on the recruited treated arm. The two causal estimands on the recruited population are identifiable via a weighting scheme that was originally designed for identifying the average treatment effect for the control and the treated individuals in the standard causal literature, respectively [Li et al., 2018]. Consequently, the corresponding overall population estimands are also identifiable. Theorem 1 becomes further interpretable under a monotonicity assumption on the enrollment status [Angrist et al., 1996].

Assumption 5. (Recruitment monotonicity) $R_{ij}(1) \geq R_{ij}(0)$ for all individuals.

Monotonicity rules out the disincentivized-recruited, so that the recruited control arm consists only of the always-recruited, and $\tau_{a,d}^O \equiv \tau_a^O$. Theorem 1 implies a weighting-based identification strategy for τ_a^O : re-weight the recruited treatment individuals by $\{1 - e(X_{ij})\}/e(X_{ij})$ and leave the recruited control individuals as they are. A similar weighting strategy allows us to identify the causal effect on the union of the always- and incentivized-recruited, $\tau_{a,c}^O$, based on the recruited sample alone.

Next, because $\tau_{a,c}^O = \nu\tau_a^O + (1 - \nu)\tau_c^O$ with $\nu = \pi_a/(\pi_a + \pi_c)$, we can identify τ_c^O as long as we can identify the relative proportions of the always- and incentivized-recruited groups. The next result shows how to identify ν and consequently τ_c^O .

Theorem 2. If Assumptions 1, 2 and 5 hold, the proportion of the always-recruited among the always- and incentivized-recruited individuals is identifiable as

$$\nu \equiv \frac{\pi_a}{\pi_a + \pi_c} = \frac{\pi^t(1 - p^t)}{(1 - \pi^t)p^t}, \quad (6)$$

where $\pi^t = \Pr(Z = 1)$ and $p^t = \Pr(Z = 1 \mid R = 1)$ is the probability of treatment in the overall and the recruited population, respectively. If Assumptions 3B and 4 also hold, the causal effect among the incentivized group is identifiable as

$$\tau_c^O = \frac{\tau_{a,c}^O - \nu\tau_a^O}{1 - \nu},$$

with $\tau_{a,c}^O, \tau_a^O$ being identified from Theorem 1.

From Lemma 1, the probability of treatment in the overall population π^t is equal to the probability of treatment for a cluster $P(Z^c = 1)$, and therefore all quantities in Theorem 2 are identifiable based on the observed data among the recruited alone. That is, Theorems 1 and 2 show that we can identify the causal effect among the always- and incentivized-recruited subpopulations of the overall population using data *only* on the recruited sample. To characterize these latent subpopulations, we can calculate the weighted average of the covariates in the always- and incentivized-recruited strata using the estimated working propensity score (introduced in Section 4). We illustrate this approach in Section 7.

4. Estimation via Working Propensity Score Weighting

4.1 Causal Effect Estimators

Given the identification formulas (3), (4) and (5), we propose the corresponding Hajék weighting estimators for τ^R , τ_a^O and $\tau_{a,c}^O$ using only the sample of recruited individuals:

$$\hat{\tau} = \frac{\sum_{ij} w_1(X_{ij}) Z_{ij} Y_{ij}}{\sum_{ij} w_1(X_{ij}) Z_{ij}} - \frac{\sum_{ij} w_0(X_{ij}) (1 - Z_{ij}) Y_{ij}}{\sum_{ij} w_0(X_{ij}) (1 - Z_{ij})}, \quad (7)$$

with the weights $\{w_0(x) = 1/(1 - e(x)), w_1(x) = 1/e(x)\}$ for τ^R , $\{w_0(x) = 1, w_1(x) = (1 - e(x))/e(x)\}$ for τ_a^O , and $\{w_0(x) = e(x)/(1 - e(x)), w_1(x) = 1\}$ for $\tau_{a,c}^O$. We estimate ν in (6) by substituting the probabilities of treatment in the overall and recruited population with the randomization probability at the cluster level $\Pr(Z^c = 1)$, and the proportion of treated individuals in the recruited sample, respectively. Then, we estimate τ_c^O using the estimators $\hat{\nu}$, $\hat{\tau}_a^O$ and $\hat{\tau}_{a,c}^O$ based on Theorem 2. We show that the resulting causal estimators are consistent and asymptotically normal using M-estimation [Van der Vaart, 2000].

Theorem 3. *If Assumptions 1, 2, 3B, 4 and 5 hold, and under standard regularity conditions, we have that $\sqrt{J}\{(\hat{\tau}_a^O, \hat{\tau}_c^O, \hat{\tau}^R)^\top - (\tau_a^O, \tau_c^O, \tau^R)^\top\} \rightarrow N(0, \Sigma)$, as $J \rightarrow \infty$ where the form of the asymptotic covariance matrix Σ is given in the Supplement.*

In brief, the regularity conditions outlined in the Supplement state that cluster sizes and outcomes are bounded quantities. Furthermore, the cluster size is assumed to be non-informative of the potential outcomes. Our results readily extend to the case of informative cluster sizes, and can be extended to address a treatment effect estimand that weighs each cluster equally [Papadogeorgou et al., 2019, Kahan et al., 2023, Wang et al., 2024].

4.2 Estimation of the working propensity score

In practice, the working propensity score $e(x)$ is usually unknown and has to be estimated. Note that the ratio of recruitment probability $\delta(x)$ defined in (2) is related to the working propensity score $e(x)$ through $e(x) = \delta(x)/\{\delta(x) + r^{-1}\}$, where $r = \Pr(Z^c = 1)/\Pr(Z^c = 0)$ (see Supplement B.4). Under recruitment monotonicity (Assumption 5), $\delta(x)$ is also related to the probability of principal stratum membership as

$$\delta(x) = \Pr(S = a \mid S \in \{a, c\}, X = x)^{-1} \geq 1$$

(see Supplement B.3). Therefore, a parametric specification can be placed on $\Pr(S = a \mid S \in \{a, c\}, X = x)$, $\delta(x)$, or $e(x)$, but it must satisfy that $\delta(x) \geq 1$. For propensity score estimation in general, the traditional approach is to specify a logistic model on $e(x)$. However, such specification is not compatible with data from experiments with recruitment bias because it does not enforce that $\delta(x) \geq 1$. Interestingly, this result on the recruitment probability ratio does not rely on the key identifying assumptions on the recruitment (Assumptions 3A and 3B) or positivity (Assumption 4), other than the fact that the function $\delta(x)$ is well-defined (Assumption 3A).

For that reason, we choose to specify $\delta(x)$ instead of $e(x)$. Specifying $\delta(x)$ also reflects the correct temporal order of the variables since the treatment occurs temporally *before* recruitment. We adopt a specification on $\Pr(S = a \mid S \in \{a, c\}, X = x)$ based on parameters α , which leads to a parametric specification on $\delta(x)$ and $e(x)$. We specify a logistic model $\Pr(S = a \mid S \in \{a, c\}, X = x) = \text{expit}(x^T \alpha)$, which implies that $\delta(x; \alpha) = 1 + \exp(-x^T \alpha)$ and $\text{logit}\{e(x; \alpha)\} = \log\{r\delta(x; \alpha)\}$.

To estimate α , it is important to recognize two features of the data structure. First, all individuals within the same cluster have the same treatment. This precludes the inclusion of cluster-level random or fixed effects in the propensity score model since the corresponding parameters would not be identifiable. Second, by definition, the working propensity score model is a marginal model with respect to clusters and not conditional on any cluster-specific intercept. Thus, a robust and computationally efficient approach for estimation is to maximize the pseudo-likelihood among the recruited sample assuming working independence: $\text{pseudo-}\mathcal{L}(\alpha) = \prod_{ij} e(X_{ij}; \alpha)^{Z_{ij}} \{1 - e(X_{ij}; \alpha)\}^{1-Z_{ij}}$, which under the specification above becomes: $\text{pseudo-}\mathcal{L}(\alpha) = \prod_{ij} \log\{r\delta(X_{ij}; \alpha)\}^{Z_{ij}} \{1 - \log\{r\delta(x; \alpha)\}\}^{1-Z_{ij}}$. A standard optimizer can be used to find the maximizer of the pseudo-likelihood $\hat{\alpha}$. We use the function `optim` in R. Since the working propensity score is primarily used to balance treated and control recruited populations rather than perfectly

predict the assignment, assuming working independence during estimation does not bias the final causal effect estimator (also empirically supported by our simulation results in Section 6). Extending the asymptotic distributions for the causal estimators in Theorem 3 to the case with the estimated propensity scores can be readily achieved by modifying the estimating function to include the pseudo-likelihood contribution for each cluster. We found in simulations (Section 6) that a bootstrap procedure that resamples clusters performs well for inference based on the true or the estimated working propensity scores. An R package for estimating the working propensity score and implementing our methods is available at <https://github.com/gpapadog/CRTrecruit>.

Finally, we note that alternative specifications of $\delta(x)$ may be considered, as long as they impose that $\delta(x) \geq 1$. This is a topic for future research.

5. Sensitivity Analysis under Non-Ignorable Recruitment

The ignorable recruitment assumption (3B) is central to identifying causal effects for the overall population. We develop a sensitivity analysis to assess this assumption within the Rosenbaum’s bounds framework [Rosenbaum, 2002]. Assume there exists an unmeasured covariate U that is necessary for Assumption 3B, that is, $R \perp\!\!\!\perp Y(0), Y(1) \mid U, X, Z$, but $R \not\perp\!\!\!\perp Y(0), Y(1) \mid X, Z$. Then, consistent estimation of the causal effects requires that we use

$$\delta^*(x, u) = \frac{\Pr(R = 1 \mid U = u, X = x, Z = 1)}{\Pr(R = 1 \mid U = u, X = x, Z = 0)},$$

and the corresponding working propensity score $e^*(x, u)$, instead of $\delta(x)$ and $e(x)$. We consider violations to the ignorable recruitment assumption with respect to U up to a factor Γ (the sensitivity parameter) as

$$\Gamma^{-1} \leq \rho(x, u) = \delta(x) / \delta^*(x, u) \leq \Gamma. \quad (8)$$

A larger value of Γ allows a greater degree of violation. To conduct a sensitivity analysis, we need to bound the chosen estimator for each fixed value of Γ . We focus on the Hajék estimator (7), which is more efficient than the Horvitz-Thompson estimator, but is technically more challenging for sensitivity analysis [Aronow and Lee, 2013, Papadogeorgou et al., 2022].

For the causal effect on the always-recruited, τ_a^O , the weights of the individuals from the control clusters are equal to 1. Therefore it suffices to focus on the part of the estimator corresponding to the individuals from the treated clusters. Denote the weights for the treated individuals in (7) for τ_a^O as $w_{1ij}^* = w_1^*(X_{ij}, U_{ij}) = \{1 - e^*(X_{ij}, U_{ij})\} / e^*(X_{ij}, U_{ij})$. The following proposition provides

an algorithm to acquire the bounds of the estimator for violations of the ignorable recruitment assumption up to Γ .

Proposition 1. *Maximizing (minimizing) $\tau_1^* = \frac{\sum_{ij} w_{1ij}^* Z_{ij} Y_{ij}}{\sum_{ij} w_{1ij}^* Z_{ij}}$ under the Γ -violation in (8) is equivalent to solving the linear program that maximizes (minimizes) $\sum_{ij} \lambda_{ij} w_{1ij} Z_{ij} Y_{ij}$ with respect to λ_{ij} subject to three constraints: (a) $\kappa\Gamma^{-1} \leq \lambda_{ij} \leq \kappa\Gamma$, (b) $\sum_{ij} \lambda_{ij} w_{1ij} Z_{ij} = 1$, and (c) $\kappa \geq 0$, where $w_{1ij} = w_1(X_{ij})$ is the weight of individual i under treatment and working propensity score $e(X_{ij})$, and κ is a parameter of the linear program.*

Bounding the causal effect on the incentivized-recruited subpopulation is more complicated because it involves a non-linear optimization problem. Instead, we acquire bounds for the causal effect estimator of the union of the always- and incentivized-recruited individuals, $\hat{\tau}_{a,c}^O$, using a procedure similar to the one in Proposition 1. Then, the bounds for the two estimators are combined according to the formula in Theorem 2 to form bounds for the estimator for the incentivized-recruited principal causal effect. This procedure is computationally efficient because it only requires solving two relatively simple linear programs. It leads to conservative bounds under the Γ -violation of the ignorability assumption in the sense that the resulting bounds are no narrower than the sharp bounds that would be obtained via directly optimizing the causal effects of the incentivized-recruited subpopulation. We provide additional details in Supplement B.8.

6. Simulation Studies

We performed simulations to study the differences between the causal estimands in the overall and recruited populations and evaluate the properties of the estimators proposed in Section 4. We considered 36 simulation scenarios, described in Table 1. In each scenario, we generated J clusters of size 100 as the overall population. The cluster-level treatment was assigned to 50% or 25% of the clusters, representing balanced and imbalanced designs, respectively. The proportion of recruited individuals depended on the specific data generating process and varied from 21% to 38%. Data under Scenario B and C had the highest and lowest proportion of recruitment, respectively. The proportion of treated individuals in the recruited population varied from 40% to 73%. Data under Scenarios A and B had the same prevalence of treatment within the recruited population. We considered three individual-level and two cluster-level covariates, and generated outcomes with treatment effect heterogeneity as well as a moderate intraclass correlation coefficient (equal to 0.1). We considered two cases corresponding to moderate and strong separation in the

covariates of the always- and incentivized recruited individuals (Case 1 & 2). The true overall average treatment effect, τ^O , was 3. The average treatment effect varied within each subpopulation across the 36 scenarios and across subpopulations, from 2.71 to 2.81 for the recruited, from 2.64 to 2.68 for the always-recruited, and from 2.88 to 3.07 for the incentivized-recruited. The average effect among the recruited varied with the treatment probability since the corresponding subpopulation changes, even though the average effect on the overall or the always-recruited populations remain constant. This illustrates that estimands on the recruited population are less interpretable, since their interpretation can be driven by the experimental design. Full details on the data generating process are given in Supplement C.1.

We evaluated the estimators in terms of bias, variance and coverage of 95% confidence intervals. We considered interval estimators based on the asymptotic theory in Theorem 3 for the known propensity score. We also consider interval estimators based on normal approximation with a nonparametric bootstrap procedure for the estimated propensity score. This bootstrap procedure resamples clusters while holding the proportion of treated clusters fixed to maintain the overall prevalence of treatment. When a cluster is sampled, all individuals in the cluster are included in the bootstrap sample.

In Figure 1, we show the true average causal effect among the different subpopulations and the estimates of the naïve estimator, under Scenario B with a balanced design. The naïve estimator, which is defined as the difference of mean outcomes among the treated recruited and the control recruited individuals, does not estimate a causal effect over any of the populations of interest across all scenarios in our simulations (see also Figure S.1 of the Supplement).

The proportion of always-recruited among the always- and incentivized-recruited in (6) is well-estimated across scenarios: Figure 2 shows the true and estimated values in our simulations, orga-

Table 1: Specifications of the data generating models for the 36 simulation scenarios. Each scenario corresponds to a different combination of principal strata prevalences, covariate differences in always and incentivized-recruited, treatment proportion, and number of clusters.

Principal strata prevalences	Scenario A	20% Always, 20% Complier, 60% Never
	Scenario B	25% Always, 25% Complier, 50% Never
	Scenario C	15% Always, 25% Complier, 60% Never
Covariate separation of always- and complier-recruited	Case 1	Strong
	Case 2	Moderate
Treatment proportion	Balanced	Probability of cluster being treated is 50%
	Imbalanced	Probability of cluster being treated is 25%
Number of clusters J	200, 500, or 800	

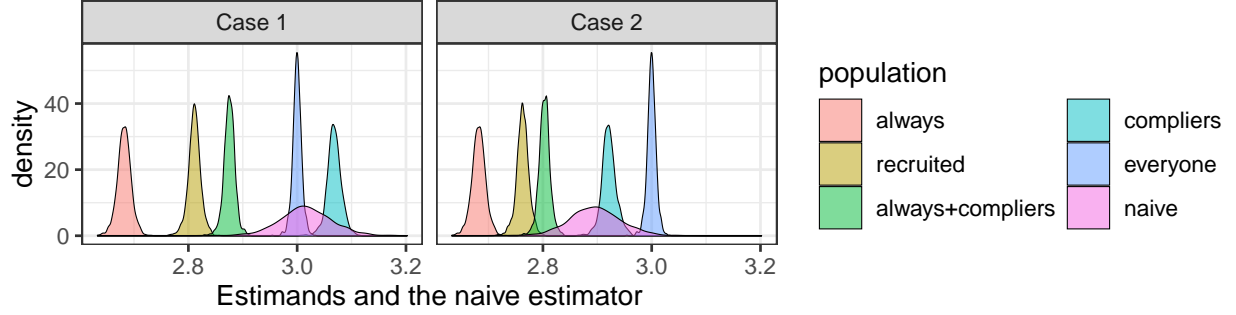


Figure 1: Distribution of sample causal effect among different populations and estimated naïve difference of mean outcomes across 500 simulated data sets and number of clusters $J \in \{200, 500, 800\}$ in Scenario B under a balanced design in Case 1 (left) and Case 2 (right).

nized by scenario and case. The vertical lines show the theoretical values for the true proportion. We observe that there is variability around the true proportion from data set to data set, but the estimation procedure is accurate, and it improves as the number of clusters increases. Furthermore, the pseudo-likelihood estimators for the working propensity score parameters are unbiased with increasing precision under a larger number of clusters (as shown in Figure S.5 of the Supplement). Figure 3 shows the difference between the estimated and the true causal effect when using the proposed estimators based on the estimated working propensity score model across 500 simulated data sets and the various simulation scenarios under a balanced design. We notice that the causal estimators for the effect on the recruited population, and the always-recruited and incentivized-recruited subpopulations are unbiased throughout, and they become more precise when the number of clusters increases. This illustrates that our approach accurately estimates the causal effect for the recruited population, and two interpretable subsets of the overall population. The causal estimators based on the true propensity score, or under imbalanced designs are also unbiased; these results are shown in the supplement.

The coverage of the asymptotic 95% intervals for the estimator that uses the known propensity score varied from 90 to 97% across all scenarios and estimands. The coverage of the bootstrap 95% intervals varied from 92.7 to 98.6% for the estimands defined on the recruited population, and the always-recruited subpopulation. Coverage of 95% intervals based on the bootstrap was lower but still reasonable for the average causal effect on the incentivized-recruited subpopulation, ranging from 85 to 93%. A figure for the coverage across simulation scenarios is shown in the supplement.

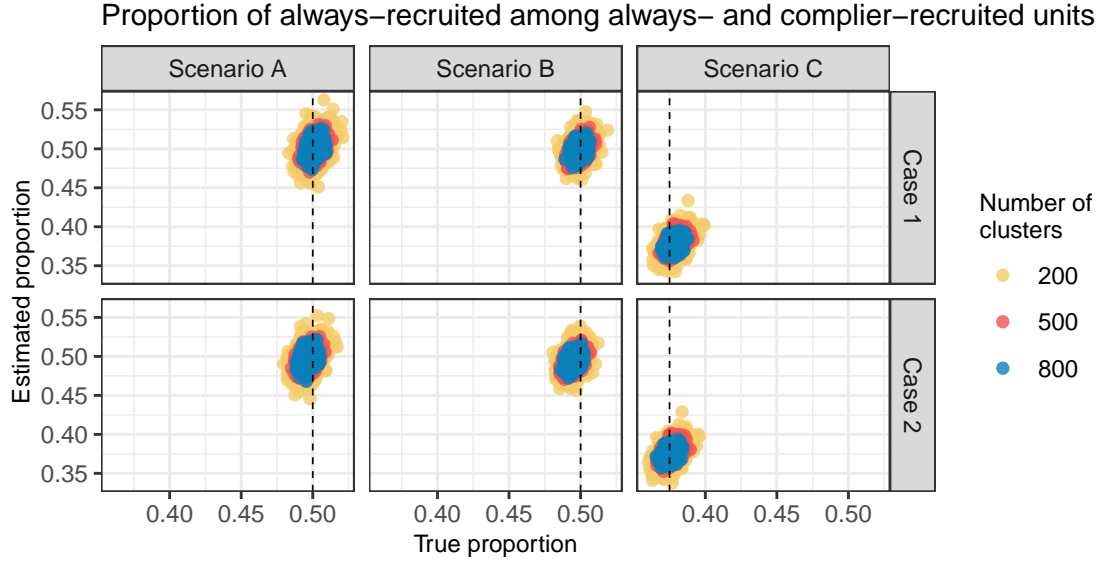


Figure 2: True sample proportion of always-recruited among the always- and complier recruited individuals (x-axis) and estimated proportion (y-axis) across 500 data sets and $J \in \{200, 500, 800\}$ number of clusters (shown in color) organized across the scenario and cases in our simulations. The vertical lines correspond to approximate theoretical values for the true super-population proportion.

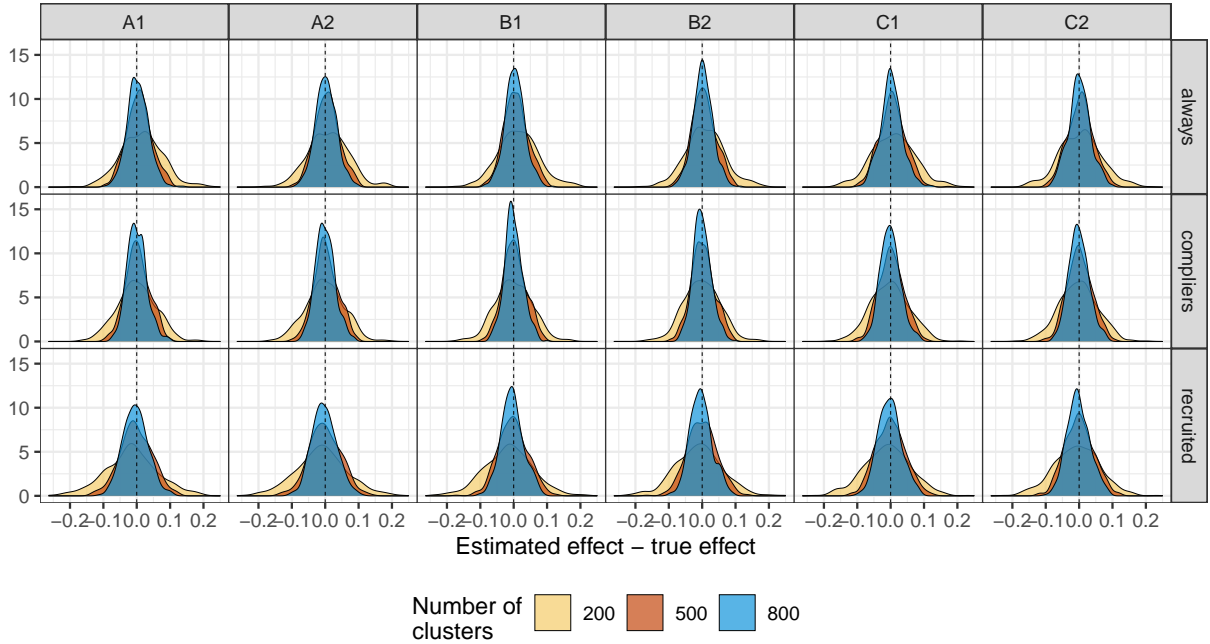


Figure 3: True minus estimated causal effect among the always-recruited, the incentivized-recruited, and the recruited populations, across 500 data sets and under the 6 scenario-case combinations, and 3 choices for the number of clusters, when the probability of cluster treatment is 0.5.

7. Application to the ARTEMIS Study

Among the original 10,976 patients, we excluded hospitals whose sample size is less than 15, resulting in a final sample size of 10,400 from 203 hospitals with 108 intervention hospitals with 6,254 patients in total and 95 control hospitals with 4,146 patients in total. We compared the covariate distribution for hospital- and patient-level covariates among the recruited, and found that a number of important covariates, such as race, education and prior P2Y₁₂ use, were not balanced between arms. Details are given in Supplement E. This observation implies potential differential recruitment due to treatment assignment. Since the intervention hospitals offered vouchers for P2Y₁₂ to the enrolled patients without out-of-pocket cost and the control hospitals offered nothing, the monotonicity assumption is considered plausible.

We focused on the outcome of binary medication persistence (equal to 1 if the patient continued on the medication) as reported by the patient at their 1-year follow-up appointment. The difference of the mean outcome among the recruited treated and control patients is 2.75% with 95% confidence interval (CI), 0.59 to 5.05%. We estimated the working propensity score model based on the pseudo-likelihood technique detailed in Section 4.2 and using all the available covariates listed in Table S.4 of the Supplement. Using the proposed method, we estimated the proportion (and 95% confidence interval) of always-recruited among the always- and incentivized-recruited individuals to be $\hat{p} = 0.754$ (95% CI: 0.56 to 1). We estimated the causal effect on the always-recruited to be 2.83% (95% CI: 0.87 to 4.75%), on the incentivized-recruited to be 2.63% (95% CI: -3.55 to 4.41%), and on the recruited population to be 2.80% (95% CI: 0.6 to 4.44%). The results imply potentially heterogeneous treatment effects across subpopulations. That is, the voucher appears to increase P2Y₁₂ medication persistence among the always-recruited patients with the associated confidence interval excluding zero. However, it may have had a weaker effect on the incentivized-recruited as the confidence interval is much wider and includes the null. Therefore, the effect in the recruited population is largely driven by the always-recruited subpopulation.

From a policy standpoint, to achieve better medication persistence, resources should be allocated to the always-recruited, who are arguably more motivated to take the medication. Unfortunately, principal stratum membership for each individual is only partially observed, rendering exact targeting of the always-recruited individuals somewhat challenging. We study the covariate distribution of the always-recruited in comparison to the recruited population to obtain additional insights into this important subpopulation. Table 2 displays the weighted mean of each covariate in

the always- and incentivized-recruited strata. Notable difference exists in some covariates between the two strata. Specifically, the always-recruited are more likely to be younger, women, non-White, have college education, have comorbidities, and not have had prior coronary artery bypass graft (CABG) surgery. Therefore, implicit targeting of the always-recruited subpopulation can be performed by targeting patients with such covariate profiles. In addition, we found that these results may be sensitive to the violations of the ignorable recruitment assumption, with minimum Γ value (for which the bounds include zero) equal to 1.13 for the always-recruited, and only 1.02 for the incentivized-recruited.

Table 2: Mean and standard deviation of the covariates among the recruited population, the always-recruited population and the incentivized-recruited subpopulation in the ARTEMIS.

Covariates	Recruited	Always Recruited	Incentivized Recruited
Age	61.74 (11.61)	61.53 (11.59)	62.35 (11.64)
Male	0.687 (0.464)	0.645 (0.479)	0.806 (0.396)
White	0.885 (0.319)	0.862 (0.345)	0.951 (0.217)
Black	0.096 (0.294)	0.114 (0.318)	0.044 (0.204)
College	0.510 (0.500)	0.536 (0.499)	0.437 (0.496)
Prior MI	0.201 (0.401)	0.201 (0.401)	0.202 (0.402)
Private insurance	0.641 (0.480)	0.646 (0.478)	0.627 (0.484)
Prior P2Y ₁₂	0.141 (0.348)	0.143 (0.350)	0.134 (0.341)
Hemoglobin Level	12.843 (2.046)	12.875 (2.032)	12.751 (2.082)
Hypertension	0.686 (0.464)	0.716 (0.451)	0.603 (0.489)
DES use	0.811 (0.391)	0.810 (0.392)	0.815 (0.389)
Diabetes	0.324 (0.468)	0.333 (0.471)	0.298 (0.458)
Employment	0.477 (0.500)	0.483 (0.500)	0.462 (0.499)
Multivessel disease	0.465 (0.499)	0.460 (0.498)	0.478 (0.500)
Prior CABG surgery	0.014 (0.118)	0.012 (0.109)	0.021 (0.142)

Furthermore, we estimated the effect on the always- and incentivized-recruited—equal to the effect on the treated recruited population—to be 2.78% (95% CI: 0.43 to 4.37%). In ARTEMIS, individuals in the treated hospitals were easily recruited because the intervention allows them to acquire their medication without co-payment. Therefore, the recruited patient population in the treated hospitals is arguably more representative of the overall population, and the effect on the treated recruited subpopulation is likely to be well representative of the effect on the overall population. Therefore, our analysis suggests that removing co-payment barriers increases the persistence of P2Y₁₂ inhibitor for patients with myocardial infarction. This conclusion is based on the presumption that recruitment in treated clusters is almost universal, something that should be evaluated conceptually separately in each study.

8. Discussion

Due to logistical and practical constraints, cluster randomized experiments are prone to post-randomization selection bias. In this paper, we established the conditions under which cluster randomization implies individual randomization, and clarified the different target populations and corresponding estimands. We provided weighting-based nonparametric identification results for these estimands based solely on the recruited sample. We also developed an interpretable sensitivity analysis method to assess the impact due to violation of the ignorable recruitment assumption.

We have focused on addressing the identification challenges due to non-random recruitment, but assumed away other potential complications that might arise in cluster randomized experiments. Importantly, we have operated under the assumption of constant cluster membership such that the treatment assignment does not affect individual choice of clusters. This assumption effectively reduces the number of possible principal strata and is plausible in scenarios where the sampling frame for each cluster is determined before the start of the study and thus the cluster membership is a pre-treatment characteristic. Schochet [2022, 2024] discussed scenarios where the treatment assignment affects cluster membership, in which case the causal estimands are redefined and stronger identification assumptions are invoked for point identification. It would be desirable to expand our work to address treatment-dependent cluster membership and sensitivity methods for these stronger assumptions. Finally, it remains an open question whether identifiability of overall effects can be achieved under weaker assumptions than the ones employed here. We found that non-differential recruitment and monotonicity (Assumptions 3A and 5) imply that $S \perp\!\!\!\perp Y(0), Y(1) \mid X, S \in \{a, c\}$ (see Proposition S.3), which is related to our identification Assumption 3B but weaker, since it involves individuals in only two of the principal strata. Some preliminary simulation studies indicate that this weaker assumption might suffice for identification (see Supplement D), and work by Schochet [2024] involves an identifying ignorability assumption that is related to the one here. Further theoretical work in this direction is warranted.

Acknowledgements

This research was supported, in part, by the Patient-Centered Outcomes Research Institute (PCORI) contract ME-2019C1-16146. The contents of this article are solely the responsibility of the authors and do not necessarily represent the view of PCORI.

Supplementary materials for
 “Addressing selection bias in cluster randomized experiments via weighting”
 by
 Georgia Papadogeorgou, Bo Liu, Fan Li, Fan Li

A. Notation

To facilitate clear presentation, we first list in Table S.1 the notation that is used throughout the manuscript and supplement.

Table S.1: Glossary of notation.

J	Total number of clusters
Z_j^c, V_j^c	Cluster level treatment and covariates
N_j, N	Number of individuals in the overall population for cluster j and in all the clusters
n_j, n	Number of recruited individuals in cluster j and in all the clusters
Z_{ij}, V_{ij}, X_{ij}	Unit-level treatment, covariates, and the vector of both unit and cluster-level covariates
$R_{ij}(0), R_{ij}(1), R_{ij}$	Potential and observed values of the unit-level recruitment
$Y_{ij}(0), Y_{ij}(1), Y_{ij}$	Potential and observed values of the outcome
τ^O, τ^R	Average effects on the overall and recruited populations
τ_z^R	Average effect on the recruited populations with treatment level z
τ_s^O	Average effects for individuals in the overall population with recruitment principal stratum $S = s$
\sum_{ij}	Represents a sum over the recruited individuals, for example $\sum_{ij} Y_{ij} = \sum_{j=1}^J \sum_{i=1}^{n_j} R_{ij} Y_{ij}$

We use $i \in j$ to denote the individuals within cluster j (recruited or not), and $i \in j^*$ to denote the recruited subjects in cluster j . Then, summations of the form $\sum_{i=1}^N$ can be re-written as $\sum_{j=1}^J \sum_{i \in j}$, and the summations of the form $\sum_{i=1}^n$ can be re-written as $\sum_{j=1}^J \sum_{i \in j^*}$.

B. Proofs

B.1 Randomized treatment on the overall population

Proof of Lemma 1. Here, it is more convenient to use the single index notation. Specifically, let $i = 1, 2, \dots, N$ denote a randomly chosen unit in the overall population of the J clusters, and let $Q_i = j$ denote that unit i belongs to cluster j . First, we start by showing that the cluster-level randomization of treatment, also implies that the cluster-level treatment is independent of cluster memberships (based on Assumption 1), formalized as:

$$\Pr(Z_j^c = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N, Q_i) = \Pr(Z_j^c = 1). \quad (\text{S.1})$$

In other words, knowing which cluster subject i belongs to does not provide any information on the cluster treatment level. Note that Assumption 1 implies that, among the overall population,

$$\Pr(Q_i = j \mid \{Z_j^c, V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) = \Pr(Q_i = j \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N).$$

Then

$$\begin{aligned} & \Pr(Z_j^c = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N, Q_i) \\ &= \frac{\Pr(Q_i \mid Z_j^c = 1, \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N)}{\Pr(\{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N, Q_i)} \times \\ & \quad \times \Pr(Z_j^c = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \Pr(\{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \\ &= \frac{\Pr(Q_i \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N)}{\Pr(\{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N, Q_i)} \times \quad (\text{From Assumption 1}) \\ & \quad \times \Pr(Z_j^c = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \Pr(\{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \\ &= \Pr(Z_j^c = 1) \quad (\text{From Assumption 2}) \end{aligned}$$

We have:

$$\begin{aligned} & \Pr(Z_i = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) = \\ &= \sum_{j=1}^J \Pr(Z_i = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N, Q_i = j) \Pr(Q_i = j \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^J \Pr(Z_j^c = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N, Q_i = j) \Pr(Q_i = j \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \\
&= \Pr(Z_j^c = 1) \sum_{j=1}^J P(Q_i = j \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) \quad (\text{From (S.1)}) \\
&= \Pr(Z_j^c = 1)
\end{aligned}$$

Further, (S.1) implies that $\Pr(Z_j^c = 1 \mid Q_i) = \Pr(Z_j^c = 1)$ and in turn

$$\begin{aligned}
\Pr(Z_i = 1) &= \sum_{j=1}^J P(Z_i = 1 \mid Q_i = j) \Pr(Q_i = j) \\
&= \sum_{j=1}^J P(Z_j^c = 1 \mid Q_i = j) \Pr(Q_i = j) \\
&= \Pr(Z_j^c = 1) \sum_{j=1}^J \Pr(Q_i = j) \quad (\text{From (S.1)}) \\
&= \Pr(Z_j^c = 1).
\end{aligned}$$

Putting these together: $\Pr(Z_i = 1 \mid \{V_j^c\}_{j=1}^J, \{V_i, Y_i(0), Y_i(1)\}_{i=1}^N) = P(Z_i = 1) = P(Z_j^c = 1)$.

□

B.2 Relationships among assumptions for post-treatment recruitment

Proposition S.1. *The non-differential recruitment Assumption 3A is equivalent to subset ignorability, $Z \perp\!\!\!\perp Y(0), Y(1) \mid X, R = 1$. The ignorable recruitment Assumption 3B is equivalent to conditional independence of principal stratum and potential outcomes $S \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X$. When Assumption 3B holds, Assumption 3A also holds, though the reverse is not true.*

Proof. First we show that Assumption 3A $\implies Z \perp\!\!\!\perp Y(0), Y(1) \mid X, R = 1$. We denote $f_{\mathbf{Y}(\cdot)}$ as the distribution of $\mathbf{Y}(\cdot) = (Y(0), Y(1))$ and $\mathbf{y}(\cdot)$ as one realization.

$$\begin{aligned}
&f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid R = 1, Z = z, X = x) \\
&= \frac{P(R = 1 \mid Z = z, X = x, \mathbf{Y}(\cdot) = \mathbf{y}(\cdot)) P(Z = z \mid X = x, \mathbf{Y}(\cdot) = \mathbf{y}(\cdot)) f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid X = x)}{P(R = 1 \mid Z = z, X = x) P(Z = z \mid X = x)} \\
&= \frac{P(R = 1 \mid Z = z, X = x, \mathbf{Y}(\cdot) = \mathbf{y}(\cdot))}{P(R = 1 \mid Z = z, X = x)} f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid X = x), \\
&\quad (\text{Randomized treatment in the overall population – See Lemma 1})
\end{aligned}$$

From Assumption 3A,

$$\begin{aligned}
P(R = 1 \mid Z = 1, X = x) &= \\
&= \int P(R = 1 \mid Z = 1, X = x, \mathbf{Y}(\cdot) = \mathbf{y}(\cdot)) f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid Z = 1, X = x) d\mathbf{y}(\cdot) \\
&= \delta(x) \int P(R = 1 \mid Z = 0, X = x, \mathbf{Y}(\cdot) = \mathbf{y}(\cdot)) f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid Z = 0, X = x) d\mathbf{y}(\cdot) \\
&\hspace{25em} (\text{Randomized treatment}) \\
&= \delta(x) P(R = 1 \mid Z = 0, X = x)
\end{aligned}$$

Therefore,

$$\begin{aligned}
f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid R = 1, Z = 1, X = x) &= \\
&= \frac{\delta(x) P(R = 1 \mid Z = 0, X = x, \mathbf{Y}(\cdot) = \mathbf{y}(\cdot))}{\delta(x) P(R = 1 \mid Z = 0, X = x)} f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid X = x) \\
&= f_{\mathbf{Y}(\cdot)}(\mathbf{y}(\cdot) \mid R = 1, Z = 0, X = x),
\end{aligned}$$

which gives us that $Z \perp\!\!\!\perp Y(0), Y(1) \mid X, R = 1$.

Next we show that $Z \perp\!\!\!\perp Y(0), Y(1) \mid X, R = 1 \implies$ Assumption 3A. Consider

$$\begin{aligned}
&\frac{P(R = 1 \mid X, Y(0), Y(1), Z = 1)}{P(R = 1 \mid X, Y(0), Y(1), Z = 0)} = \\
&= \frac{P(Z = 1 \mid X, Y(0), Y(1), R = 1) P(R = 1 \mid X, Y(0), Y(1))}{P(Z = 0 \mid X, Y(0), Y(1), R = 1) P(R = 1 \mid X, Y(0), Y(1))} \bigg/ \frac{P(Z = 1 \mid X, Y(0), Y(1))}{P(Z = 0 \mid X, Y(0), Y(1))} \\
&= \frac{P(Z = 1 \mid X, Y(0), Y(1), R = 1) P(Z = 0 \mid X, Y(0), Y(1))}{P(Z = 0 \mid X, Y(0), Y(1), R = 1) P(Z = 1 \mid X, Y(0), Y(1))} \\
&= \frac{P(Z = 1 \mid X, R = 1) P(Z = 0)}{P(Z = 0 \mid X, R = 1) P(Z = 1)},
\end{aligned}$$

where the last equation holds from subset ignorability for the first ratio, and from Lemma 1 for the second ratio. Therefore, the ratio of enrollment probabilities is only a function of the covariates and Assumption 3A is satisfied.

Next we show that Assumption 3B $\implies S \perp\!\!\!\perp \{Y(0), Y(1)\} \mid Z, X$. Notice that $\mathbb{I}(S \in \{a, c\}) = ZR$, $\mathbb{I}(S \in \{a, d\}) = (1 - Z)R$. Therefore, $R \perp\!\!\!\perp \{Y(0), Y(1)\} \mid Z, X$ implies $(\mathbb{I}(S \in \{a, c\}), \mathbb{I}(S \in \{a, d\})) \perp\!\!\!\perp \{Y(0), Y(1)\} \mid Z, X$. The entries of $(\mathbb{I}(S \in \{a, c\}), \mathbb{I}(S \in \{a, d\}))$ are a 1-1 map with the

stratum membership S since

$$(\mathbb{I}(S \in \{a, c\}), \mathbb{I}(S \in \{a, d\})) = \begin{cases} (0, 0) & \iff S = n \\ (1, 0) & \iff S = c \\ (0, 1) & \iff S = d \\ (1, 1) & \iff S = a. \end{cases}$$

Therefore, we have that $S \perp\!\!\!\perp \{Y(0), Y(1)\} \mid Z, X$. Both S and $\{Y(0), Y(1)\}$ are independent of Z , and hence we may drop Z from the conditioning set.

Since all the steps also work backwards, we have that $S \perp\!\!\!\perp \{Y(0), Y(1)\} \mid Z, X \implies$ Assumption 3B, which completes the proof that the two assumptions are equivalent.

Next we show that Assumption 3B implies Assumption 3A. From Assumption 3B we have that

$$\frac{P(R = 1 \mid X, Y(0), Y(1), Z = 1)}{P(R = 1 \mid X, Y(0), Y(1), Z = 0)} = \frac{P(R = 1 \mid X, Z = 1)}{P(R = 1 \mid X, Z = 0)}$$

is only a function of covariates and Assumption 3A is satisfied.

The reverse is not true, and Assumption 3A does not imply Assumption 3B. The following example is a situation where Assumption 3A holds, and Assumption 3B does not. For a binary outcome, assume that

$$P(R = 1 \mid X, Y(0), Y(1), Z) = \begin{cases} 0.5 & \text{if } Y(0) = 0 \text{ and } Z = 1, \\ 0.2 & \text{if } Y(0) = 0 \text{ and } Z = 0, \\ 0.75 & \text{if } Y(0) = 1 \text{ and } Z = 1, \\ 0.3 & \text{if } Y(0) = 1 \text{ and } Z = 0. \end{cases}$$

□

B.3 Additional information on the non-differential recruitment functional

In Proposition S.2, we establish equivalent forms of the ratio of recruitment probabilities defined in equation (2) of the manuscript.

Proposition S.2. *Under constant cluster membership (Assumption 1) and treatment randomiza-*

tion at the cluster level (Assumption 2), the ratio of recruitment probabilities in (2) is equal to

$$\frac{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 0)} = \frac{\Pr(S \in \{a, c\} \mid Y(1) = y_1, Y(0) = y_0, X = x)}{\Pr(S \in \{a, d\} \mid Y(1) = y_1, Y(0) = y_0, X = x)},$$

for all values of x, y_0, y_1 . Furthermore, if monotonicity of the recruitment indicator (Assumption 5) holds, then this ratio is equal to $1/\Pr(S = a \mid S \in \{a, c\}, Y(1) = y_1, Y(0) = y_0, X = x) \geq 1$.

Proof. Note that

$$\begin{aligned} & \Pr(R = 1 \mid Y(1), Y(0), X, Z = 1) \\ &= \sum_{s \in \{a, c, d, n\}} \Pr(R = 1 \mid Y(1), Y(0), X, Z = 1, S = s) \Pr(S = s \mid Y(1), Y(0), X, Z = 1) \\ &= \sum_{s \in \{a, c\}} \Pr(S = s \mid Y(1), Y(0), X) \quad (\text{From Lemma 1, we can drop } Z \text{ from the conditioning set.}) \\ &= \Pr(S \in \{a, c\} \mid Y(1), Y(0), X). \end{aligned}$$

Similarly,

$$\Pr(R = 1 \mid Y(1), Y(0), X, Z = 0) = \Pr(S \in \{a, d\} \mid Y(1), Y(0), X).$$

Under Assumption 5, the d stratum does not exist, and

$$\Pr(R = 1 \mid Y(1), Y(0), X, Z = 0) = \Pr(S = a \mid Y(1), Y(0), X).$$

The proof is immediate since

$$\Pr(S = a \mid S \in \{a, c\}, Y(1), Y(0), X) = \frac{\Pr(S = a \mid Y(1), Y(0), X)}{\Pr(S \in \{a, c\} \mid Y(1), Y(0), X)},$$

and since any probability is between 0 and 1, we have that $1/\Pr(S = a \mid S \in \{a, c\}, Y(1) = y_1, Y(0) = y_0, X = x) \geq 1$. \square

Proposition S.3. *Under constant cluster membership (Assumption 1) and treatment randomization at the cluster level (Assumption 2), if non-differential recruitment (Assumption 3A) and monotonicity (Assumption 5) hold, then it holds that $S_{ij} \perp\!\!\!\perp Y_{ij}(0), Y_{ij}(1) \mid X_{ij}, S_{ij} \in \{a, c\}$.*

Proof. Since

$$\delta(x) = \frac{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 0)},$$

we know from Proposition S.2, and using monotonicity, that

$$\delta(x) = 1 / \Pr(S_{ij} = a \mid S_{ij} \in \{a, c\}, Y_{ij}(1) = y_1, Y_{ij}(0) = y_0, X_{ij} = x).$$

Then, we have that

$$\begin{aligned} P(S_{ij} = a \mid X_{ij}, S_{ij} \in \{a, c\}) &= \\ &= \int_{y_0, y_1} P(S_{ij} = a \mid X_{ij}, Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, S_{ij} \in \{a, c\}) f(y_0, y_1 \mid X_{ij}, S_{ij} \in \{a, c\}) \\ &= \int_{y_0, y_1} \delta(x)^{-1} f(y_0, y_1 \mid X_{ij}, S_{ij} \in \{a, c\}) \\ &= \delta(x)^{-1}, \end{aligned}$$

assuming for simplicity that the potential outcome distribution admits a density. So we have that

$$\Pr(S_{ij} = a \mid S_{ij} \in \{a, c\}, Y_{ij}(1) = y_1, Y_{ij}(0) = y_0, X_{ij} = x) = \Pr(S_{ij} = a \mid S_{ij} \in \{a, c\}, X_{ij} = x).$$

Since S_{ij} is binary within the group $S_{ij} \in \{a, c\}$, this implies that $S_{ij} \perp\!\!\!\perp Y_{ij}(0), Y_{ij}(1) \mid X_{ij}, S_{ij} \in \{a, c\}$. \square

The function $\delta(x)$ in Assumption 3A is defined as

$$\delta(x) = \frac{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 0)},$$

and its form under assumptions with respect to principal stratum membership was given in Proposition S.2. We note here that the definition of $\delta(x)$ includes the values of the potential outcomes, and $\delta(x)$ need *not* necessarily be equal to the ratio of

$$\frac{\Pr(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 0)},$$

or functions similar to those in Proposition S.2 excluding potential outcomes. The following proposition establishes that under *additional* assumptions, this equivalence holds.

Proposition S.4. Assume that the conditions of Proposition S.2 hold. Then,

(a) If Assumption 3B holds, then $\delta(x) = \Pr(S_{ij} \in \{a, c\} \mid X_{ij} = x) / \Pr(S_{ij} \in \{a, d\} \mid X_{ij} = x)$.

(b) If Assumption 5 holds, then $\delta(x) = 1 / P(S_{ij} = a \mid S_{ij} \in \{a, c\}, X_{ij} = x)$.

$$\text{Under either assumption, } \delta(x) = \frac{\Pr(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 0)}.$$

Proof.

(a)

$$\begin{aligned} \delta(x) &= \frac{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid Y_{ij}(0) = y_0, Y_{ij}(1) = y_1, X_{ij} = x, Z_{ij} = 0)} \\ &= \frac{\Pr(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 1)}{\Pr(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 0)} \\ &\quad \text{(Using Assumption 3B for the numerator and denominator)} \\ &= \frac{P(S_{ij} \in \{a, c\} \mid X_{ij} = x, Z_{ij} = 1)}{P(S_{ij} \in \{a, d\} \mid X_{ij} = x, Z_{ij} = 0)} \\ &= \frac{P(S_{ij} \in \{a, c\} \mid X_{ij} = x)}{P(S_{ij} \in \{a, d\} \mid X_{ij} = x)} \quad \text{(From Lemma 1)} \end{aligned}$$

(Note that the second equation above is one of the results we are showing.)

(b)

$$\begin{aligned} P(S_{ij} = a \mid S_{ij} \in \{a, c\}, X_{ij} = x) &= \\ &= \int_{y_0, y_1} P(S_{ij} = a \mid S_{ij} \in \{a, c\}, X_{ij} = x) f(y_0, y_1 \mid S_{ij} \in \{a, c\}, X_{ij} = x) d(y_0, y_1) \\ &= \delta(x)^{-1} \int_{y_0, y_1} f(y_0, y_1 \mid S_{ij} \in \{a, c\}, X_{ij} = x) d(y_0, y_1) \quad \text{(From Proposition S.2)} \\ &= \delta(x)^{-1} \end{aligned}$$

Therefore, we can also write

$$\begin{aligned} \delta(x) &= \frac{1}{P(S_{ij} = a \mid S_{ij} \in \{a, c\}, X_{ij} = x)} \\ &= \frac{P(S_{ij} \in \{a, c\} \mid X_{ij} = x)}{P(S_{ij} = a \mid X_{ij} = x)} \\ &= \frac{P(S_{ij} \in \{a, c\} \mid X_{ij} = x, Z_{ij} = 1)}{P(S_{ij} = a \mid X_{ij} = x, Z_{ij} = 0)} \quad \text{(From Lemma 1)} \\ &= \frac{P(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 1)}{P(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 0)} \quad \text{(Using monotonicity)} \end{aligned}$$

□

The results in Proposition S.2 and Proposition S.4 imply that under standard conditions (constant cluster membership and randomization of the treatment assignment) and monotonicity, the function $\delta(x)$ satisfies that

$$\delta(x) = \frac{P(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 1)}{P(R_{ij} = 1 \mid X_{ij} = x, Z_{ij} = 0)} \geq 1.$$

B.4 The working propensity score

We show that enrollment positivity leads to working propensity score positivity.

Proposition S.5 (Working propensity score positivity). *If Assumption 4 holds, then there exists $\delta' > 0$ such that $\delta' < e(x) < 1 - \delta'$ for all x .*

Proof. Remember that

$$e(x) = \frac{P(R = 1 \mid Z = 1, X = x)}{P(R = 1 \mid X = x)} P(Z = 1). \quad (\text{S.2})$$

From Assumption 4 we have that $P(R = 1 \mid Z = z, X = x) > \delta$, which gives us that $e(x) > \delta P(Z = 1)$. It also implies that $P(R = 1 \mid X = x) > \delta$, which in turn gives us that $e(x) < P(Z = 1)/\delta$. By setting $\delta' = \min\{\delta P(Z = 1), 1 - P(Z = 1)/\delta\}$, we have the result. □

Next, we show the relationship between the working propensity score $e(x)$ and the ratio of recruitment probabilities $\delta(x)$ in Assumption 3A.

Proposition S.6. *If the assumptions of constant cluster membership (Assumption 1), treatment randomization (Assumption 2), and non-differential recruitment (Assumption 3A) hold, then $e(x) = \delta(x)/\{\delta(x) + r^{-1}\}$ where $r = \Pr(Z^c = 1)/\Pr(Z^c = 0)$.*

Proof. In Proposition S.1, it was established that Assumption 3A is equivalent to $Z_{ij} \perp\!\!\!\perp \{Y_{ij}(0), Y_{ij}(1)\} \mid X_{ij}, R_{ij} = 1$. Therefore,

$$\begin{aligned} e(x) &= P(Z = 1 \mid X = x, R = 1) \\ &= P(Z = 1 \mid X = x, R = 1, Y(0), Y(1)) \\ &= \frac{P(R = 1 \mid X = x, Y(0), Y(1), Z = 1) P(Z = 1 \mid X = x, Y(0), Y(1))}{P(R = 1 \mid X = x, Y(0), Y(1))} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{1 + \frac{P(R=1 \mid X=x, Y(0), Y(1), Z=0)P(Z=0 \mid X=x, Y(0), Y(1))}{P(R=1 \mid X=x, Y(0), Y(1), Z=1)P(Z=1 \mid X=x, Y(0), Y(1))}} \\
&= \frac{1}{1 + \delta^{-1}(x)r^{-1}} \quad (\text{Using Lemma 1}) \\
&= \frac{\delta(x)}{\delta(x) + r^{-1}}
\end{aligned}$$

□

B.5 Identifiability of average causal effects for the recruited population

Here, we use the equivalence between weighting and conditioning estimands for our proof. The weighted average causal effect among a subpopulation of the recruited population with covariate distribution $g^R(x)$ is defined as

$$\tau_g^R = E[h^R(X)[Y(1) - Y(0)] \mid R = 1],$$

where $h^R(x) = g^R(x)/f^R(x)$, and $f^R(x)$ is the covariate distribution among the recruited individuals. For $g^R(x) = f^R(x)$, we have that $\tau_g^R = \tau^R$. Also, for $g^R(x) = f(x \mid Z = 1, R = 1)$ or $g^R(x) = f(x \mid Z = 0, R = 1)$, the estimand τ_g^R reverts to the average causal effect among the treated recruited individuals τ_1^R or the control recruited individuals τ_0^R , respectively.

Proposition S.7. *If Assumption 1, Assumption 2, Assumption 3A, and Assumption 4 hold, then*

$$\tau_g^R = E \left[h^R(X) \left\{ \frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)} \right\} \mid R = 1 \right],$$

for any subpopulation of the recruited population with covariate distribution $g^R(x)$ and corresponding tilting function $h^R(x)$.

Proof. We have

$$\begin{aligned}
E \left\{ h^R(X) \frac{ZY}{e(X)} \mid R = 1 \right\} &= E \left\{ E \left[h^R(X) \frac{ZY(1)}{e(X)} \mid X, Y(1), Y(0), R = 1 \right] \mid R = 1 \right\} \\
&= E \left\{ \frac{h^R(X)}{e(X)} Y(1) E[Z \mid X, Y(1), Y(0), R = 1] \mid R = 1 \right\} \\
&= E \left\{ \frac{h^R(X)}{e(X)} Y(1) P(Z = 1 \mid X, Y(1), Y(0), R = 1) \mid R = 1 \right\}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left\{ \frac{h^R(X)}{e(X)} Y(1) P(Z = 1 \mid X, R = 1) \mid R = 1 \right\} \\
&\quad \text{(From Proposition S.1 \& using Assumption 3A)} \\
&= \mathbb{E} \{ h^R(X) Y(1) \mid R = 1 \}
\end{aligned}$$

and similarly we can get that

$$\mathbb{E} \left\{ h^R(X) \frac{(1-Z)Y}{1-e(X)} \mid R = 1 \right\} = \mathbb{E} \{ h^R(X) Y(0) \mid R = 1 \},$$

which verifies the weighting-based identification formula. □

B.6 Identifiability of average causal effects for the overall population

We similarly use weighting definitions of estimands for the overall population for our proofs. The weighted average causal effect among a subpopulation of the overall population with covariate distribution $g^O(x)$ is defined as

$$\tau_g^O = \mathbb{E} [h^O(X)[Y(1) - Y(0)]],$$

where $h^O(x) = g^O(x)/f_X(x)$, and $f_X(x)$ is the covariate distribution in the overall population. For $g^O(x) = f_X(x)$, we have that $\tau_g^O = \tau^O$. Also, for $g^O(x) = f(x \mid S \in \{a, d\})$, the estimand τ_g^O reverts to the average causal effect among the always- and disincentivized-recruited individuals $\tau_{a,d}^O$, and similarly for the other principal strata.

Proof of Theorem 1. First, we establish how probabilistic statements about the recruitment indicator R given treatment can be linked to probabilistic statements about the potential recruitment value $R(z)$:

$$\begin{aligned}
\mathbb{P}(R = 1 \mid Z = z, X = x) &= \mathbb{P}(R(z) = 1 \mid Z = z, X = x) \\
&= \frac{\mathbb{P}(Z = z \mid R(z) = 1, X = x)}{\mathbb{P}(Z = z \mid X = x)} \mathbb{P}(R(z) = 1 \mid X = x) \quad (\text{S.3}) \\
&= \mathbb{P}(R(z) = 1 \mid X = x),
\end{aligned}$$

using consistency of the potential recruitment indicators and that the treatment is randomized

from Lemma 1. Similarly

$$P(R = 1 \mid Z = z) = P(R(z) = 1).$$

For the following, we will also use the result from Proposition S.1 which shows that Assumption 3B implies subset ignorability, $Z \perp\!\!\!\perp Y(0), Y(1) \mid X, R = 1$.

- (a) In this part, the subpopulation of the recruited individuals that is of interest corresponds to that with covariate density g^R of the control individuals among the recruited population, $g^R(x) = f(x \mid Z = 0, R = 1)$. The corresponding tilting function on the recruited population is denoted as $h^R(x)$. This implies that $\tau_g^R = \tau_0^R$. Also, the covariate density of interest on the overall population, g^O , corresponds to the covariate density of the always and defying-recruiters, $g^O(x) = f(x \mid S \in \{a, d\})$, with corresponding overall tilting function $h^O(x)$. Then $\tau_g^O = \tau_{a,d}^O$. First, we show that these two covariate distributions are the same.

$$\begin{aligned} g^R(x) &= f_X(x \mid Z = 0, R = 1) \\ &= \frac{P(R = 1 \mid X = x, Z = 0)P(Z = 0 \mid X = x)}{P(R = 1 \mid Z = 0)P(Z = 0)} f_X(x) \\ &= \frac{P(R(0) = 1 \mid X = x)}{P(R(0) = 1)} f_X(x) && \text{(From equation (S.3) and Lemma 1)} \\ &= f_X(x \mid R(0) = 1) \\ &= f_X(x \mid S \in \{a, d\}) \\ &= g^O(x) \end{aligned} \tag{S.4}$$

Then, we show that $E[h^R(X)Y(z) \mid R = 1] = E[h^O(X)Y(z)]$, which implies that the estimands of these two populations are equal, in that

$$\tau_g^R = E[h^R(X)(Y(1) - Y(0)) \mid R = 1] = E[h^O(X)(Y(1) - Y(0))] = \tau_g^O$$

To show this, consider the expected outcome in the tilted recruited population. Remember that this tilted population with tilting function h^R represents the recruited controls:

$$\begin{aligned} E[h^R(X) Y(1) \mid R = 1] &= \int_x \int_y y f(Y(1) = y \mid X = x, R = 1) dy h^R(x) f(x \mid R = 1) dx \\ &= \int_x \int_y y f(Y(1) = y \mid X = x, R = 1) dy g^R(x) dx \end{aligned}$$

$$\begin{aligned}
&= \int_x \int_y y f(Y(1) = y \mid Z = 1, X = x, R = 1) dy g^O(x) dx \\
&\quad \text{(From subset ignorability and the result above)} \\
&= \int_x \int_y y f(Y(1) = y \mid Z = 1, X = x) dy h^O(x) f_X^O(x) dx \\
&\quad \text{(From Assumption 3B)} \\
&= \int_x \int_y y f(Y(1) = y \mid X = x) dy h^O(x) f_X^O(x) dx \quad \text{(From Lemma 1)} \\
&= E[h^O(X)Y(1)].
\end{aligned}$$

We can similarly show that $E[h^R(X)Y(0) \mid R = 1] = E[h^O(X)Y(0)]$. Therefore, $\tau_0^R = \tau_g^R = \tau_g^O = \tau_{a,d}^O$.

(b) As long as we show that $g^R(x) = g^O(x)$ for $g^R(x) = f(x \mid Z = 1, R = 1)$ and $g^O(x) = f(x \mid S \in \{a, c\})$, then the proof will follow similarly to the steps for the first part. Here

$$\begin{aligned}
g^R(x) &= f_X(x \mid Z = 1, R = 1) \\
&= \frac{P(R = 1 \mid X = x, Z = 1)P(Z = 1 \mid X = x)}{P(R = 1 \mid Z = 1)P(Z = 1)} f_X(x) \\
&= \frac{P(R(1) = 1 \mid X = x)}{P(R(1) = 1)} f_X(x) \quad \text{(From equation (S.3) and Lemma 1)} \\
&= f_X(x \mid R(1) = 1) \\
&= f_X(x \mid S \in \{a, c\}) \\
&= g^O(x)
\end{aligned}$$

The identification formulas for $\tau_{a,d}^O$ and $\tau_{a,c}^O$ are straightforward to acquire based on the formulas for the formulas for τ_0^R and τ_1^R in Proposition S.7. \square

Proof of Theorem 2.

$$\begin{aligned}
p^t &= P(Z = 1 \mid R = 1) \\
&= \frac{P(R = 1 \mid Z = 1)P(Z = 1)}{P(R = 1 \mid Z = 1)P(Z = 1) + P(R = 1 \mid Z = 0)P(Z = 0)} \\
&= \frac{P(S \in \{a, c\} \mid Z = 1)\pi^t}{P(S \in \{a, c\} \mid Z = 1)\pi^t + P(S = a \mid Z = 0)(1 - \pi^t)} \\
&= \frac{(\pi_a + \pi_c)\pi^t}{(\pi_a + \pi_c)\pi^t + \pi_a(1 - \pi^t)},
\end{aligned}$$

where the second equation using monotonicity and Lemma 1, and the third equation uses the definitions of $\pi_{a,c}, \pi_a$ and Lemma 1 that states that the treatment is independent of the stratum memberships. Solving for $\pi_a/(\pi_a + \pi_c)$ gives us the result. Based on this and Theorem 1, τ_c^O is also identifiable. \square

B.7 Asymptotic normality of causal effect estimators

We establish the value of $1 - e(x)$ which will be useful moving forward. Note that

$$\begin{aligned} e(x) &= \frac{P(R = 1 \mid Z = 1, X = x)}{P(R = 1 \mid X = x)} P(Z = 1) \\ \iff 1 - e(x) &= \frac{P(R = 1 \mid X = x) - P(R = 1 \mid Z = 1, X = x)P(Z = 1 \mid X = x)}{P(R = 1 \mid X = x)} \\ &= \frac{P(R = 1 \mid Z = 0, X = x)}{P(R = 1 \mid X = x)} P(Z = 0) \end{aligned} \quad (\text{S.5})$$

Assumption S.1 (Regularity assumptions for asymptotic normality).

- (a) N_j is independent of the cluster average potential outcome, i.e., if the cluster average potential outcome is defined as $\bar{Y}_{j,h}^O = \frac{1}{N_j} \sum_{i \in j} Y_{ij}(1)h^O(X_{ij})$ for some tilting function $h^O(x)$, then we have that $N_j \perp\!\!\!\perp \bar{Y}_{j,h}^O$,
- (b) there exists M_N such that $P(N_j < M_N) = 1$,
- (c) there exists M_Y such that $P(|Y_{ij}| < M_Y) = 1$,

Proof of Theorem 3. To show asymptotic normality of our estimator we will use Theorem 5.41 of Van der Vaart [2000]. For this proof, we take the working propensity score and the proportion of always-recruited among the always- and incentivized-recruited groups as known. The extension to estimated propensity score and estimated proportion of always-recruited is straightforward, and discussed immediately after the proof. Since we assume monotonicity, the stratum of disincentivized-recruited individuals is empty, and quantities that correspond to $S \in \{a, d\}$ are the same as those for $S = a$.

The estimating equations Let $w_z^a(x)$ denote the weight for a unit with $Z = z$ and $X = x$ for the estimand τ_a^O , $w_z^{a,c}(x)$ the weight for a unit with $Z = z$ and $X = x$ for the estimand $\tau_{a,c}^O$, and $w_z^R(x)$ the weight for a unit with $Z = z$ and $X = x$ for the estimand τ^R .

Let $\boldsymbol{\theta}^a = (\theta_1^a, \theta_2^a, \theta_3^a, \theta_4^a)^\top$ and similarly defined vectors of length four for $\boldsymbol{\theta}^{a,c}$ and $\boldsymbol{\theta}^R$. Then, let

$$\psi^a(O_j; \boldsymbol{\theta}^a) = \begin{pmatrix} \sum_{i \in j} R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij} - \theta_1 \sum_{i \in j} R_{ij} \\ \sum_{i \in j} R_{ij} w_0^a(X_{ij}) (1 - Z_{ij}) Y_{ij} - \theta_2 \sum_{i \in j} R_{ij} \\ \sum_{i \in j} R_{ij} w_1^a(X_{ij}) Z_{ij} - \theta_3 \sum_{i \in j} R_{ij} \\ \sum_{i \in j} R_{ij} w_0^a(X_{ij}) (1 - Z_{ij}) - \theta_4 \sum_{i \in j} R_{ij} \end{pmatrix}, \quad (\text{S.6})$$

and similarly defined $\psi^{a,c}(O_j; \boldsymbol{\theta}^{a,c})$ and $\psi^R(O_j; \boldsymbol{\theta}^R)$. Then, for $\boldsymbol{\theta} = (\boldsymbol{\theta}^{a^\top}, \boldsymbol{\theta}^{a,c^\top}, \boldsymbol{\theta}^{R^\top})^\top$, our estimating equations are of length 12 and defined as

$$\psi(O_j; \boldsymbol{\theta}) = \begin{pmatrix} \psi^a(O_j; \boldsymbol{\theta}^a) \\ \psi^{a,c}(O_j; \boldsymbol{\theta}^{a,c}) \\ \psi^R(O_j; \boldsymbol{\theta}^R) \end{pmatrix}.$$

The true solution to the estimating equations We use $\psi_k(O_j; \boldsymbol{\theta})$ to denote the k^{th} entry in the vector $\psi(O_j; \boldsymbol{\theta})$. Then, for $\psi_1(O_j; \boldsymbol{\theta})$, it holds that

$$\begin{aligned} & \text{E} \left[\sum_{i \in j} R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij} - \theta_1^a \sum_{i \in j} R_{ij} \right] = 0 \\ \iff & \text{E} \left\{ \sum_{i \in j} \text{E} [R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij} \mid n_j] \right\} = \text{E}[n_j] \theta_1^a \end{aligned} \quad (\text{S.7})$$

and

$$\begin{aligned} & \text{E} [R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij} \mid n_j] \\ &= \text{E} [R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij}(1) \mid n_j] \\ &= \text{E} [R_{ij} w_1^a(X_{ij}) Y_{ij}(1) \mid Z_{ij} = 1, n_j] P(Z_{ij} = 1 \mid n_j) \\ &= \text{E} \{ \text{E} [R_{ij} w_1^a(X_{ij}) Y_{ij}(1) \mid Z_{ij} = 1, Y_{ij}(0), Y_{ij}(1), X_{ij}] \mid Z_{ij} = 1, n_j \} P(Z_{ij} = 1) \\ & \quad (X_{ij} \text{ includes cluster-level information such as } n_j) \\ &= \text{E} \{ w_1^a(X_{ij}) Y_{ij}(1) P(R_{ij} = 1 \mid Z_{ij} = 1, Y_{ij}(0), Y_{ij}(1), X_{ij}) \mid Z_{ij} = 1, n_j \} P(Z_{ij} = 1) \\ &= \text{E} \{ w_1^a(X_{ij}) Y_{ij}(1) P(R_{ij} = 1 \mid Z_{ij} = 1, X_{ij}) \mid Z_{ij} = 1, n_j \} P(Z_{ij} = 1) \quad (\text{From Assumption 3B}) \end{aligned}$$

Then, from (S.5), we have that

$$\begin{aligned}
w_1^a(X_{ij})P(R_{ij} = 1 \mid Z_{ij} = 1, X_{ij}) &= \\
&= \frac{1 - e(X_{ij})}{e(X_{ij})}P(R_{ij} = 1 \mid Z_{ij} = 1, X_{ij}) \\
&= \frac{P(Z_{ij} = 0 \mid X_{ij}, R_{ij} = 1)}{\frac{P(R_{ij}=1 \mid Z_{ij}=1, X_{ij})}{P(R_{ij}=1 \mid X_{ij})}P(Z_{ij} = 1)}P(R_{ij} = 1 \mid Z_{ij} = 1, X_{ij}) \\
&= \frac{P(Z_{ij} = 0 \mid X_{ij}, R_{ij} = 1)P(R_{ij} = 1 \mid X_{ij})}{P(Z_{ij} = 1)} \\
&= \frac{P(R_{ij} = 1 \mid Z_{ij} = 0, X_{ij})P(Z_{ij} = 0)}{P(Z_{ij} = 1)} \\
&= \frac{P(R_{ij}(0) = 1 \mid X_{ij})P(Z_{ij} = 0)}{P(Z_{ij} = 1)}. \tag{From (S.3)}
\end{aligned}$$

Returning to the previous quantity we have that

$$\begin{aligned}
E\{R_{ij}w_1^a(X_{ij})Z_{ij}Y_{ij} \mid n_j\} &= \\
&= E\{Y_{ij}(1)P(R_{ij}(0) = 1 \mid X_{ij}) \mid n_j\} \frac{P(Z_{ij} = 0)}{P(Z_{ij} = 1)}P(Z_{ij} = 1) \\
&= P(Z_{ij} = 0)E\{Y_{ij}(1)h_a^O(X_{ij}) \mid n_j\}. \quad (\text{for } g_a^O(x) = f_X(x \mid S = a) \text{ based on (S.4)})
\end{aligned}$$

Returning back to (S.7), we have that

$$\begin{aligned}
E\left\{\sum_{i \in j} E[R_{ij}w_1^a(X_{ij})Z_{ij}Y_{ij} \mid n_j]\right\} &= P(Z_{ij} = 0) E\left\{n_j E\left[\frac{1}{n_j} \sum_{i \in j} Y_{ij}(1)h_a^O(X_{ij}) \mid n_j\right]\right\} \\
&= P(Z_{ij} = 0)E(n_j)E\left[\frac{1}{n_j} \sum_{i \in j} Y_{ij}(1)h_a^O(X_{ij})\right] \\
&\tag{From Assumption S.1(a)} \\
&= P(Z_{ij} = 0)E(n_j)E[Y_{ij}(1)h_a^O(X_{ij})].
\end{aligned}$$

Following similar steps for the remaining θ s while noting that $E[h_a^O(X_{ij})] = 1$, we have that the vector $\theta_0^a = (\theta_{01}^a, \theta_{02}^a, \theta_{03}^a, \theta_{04}^a)^\top$ for

$$\begin{aligned}
\theta_{01}^a &= P(Z_{ij} = 0)E[Y_{ij}(1)h_a^O(X_{ij})] \\
\theta_{02}^a &= P(Z_{ij} = 0)E[Y_{ij}(0)h_a^O(X_{ij})]
\end{aligned}$$

$$\theta_{03}^a = P(Z_{ij} = 0)$$

$$\theta_{04}^a = P(Z_{ij} = 0)$$

satisfies that $E\{\psi^a(O_j; \theta_0^a)\} = 0$. Similarly defined $\theta_0^{a,c}$ and θ_0^R satisfy that $E\{\psi^{a,c}(O_j; \theta_0^{a,c})\} = 0$ and $E\{\psi^R(O_j; \theta_0^R)\} = 0$, and as a result $\theta_0 = (\theta_0^{a\top}, \theta_0^{a,c\top}, \theta_0^{R\top})^\top$ satisfy that $E\{\psi(O_j; \theta_0)\} = 0$. Furthermore, the causal effects of interest can be written as

$$\tau_a^O = \frac{\theta_{01}^a}{\theta_{03}^a} - \frac{\theta_{02}^a}{\theta_{04}^a}, \quad \tau_{a,c}^O = \frac{\theta_{01}^{a,c}}{\theta_{03}^{a,c}} - \frac{\theta_{02}^{a,c}}{\theta_{04}^{a,c}}, \quad \text{and} \quad \tau^R = \frac{\theta_{01}^R}{\theta_{03}^R} - \frac{\theta_{02}^R}{\theta_{04}^R}.$$

The empirical solution to the estimating equations Next we shift our attention to the empirical version of the estimating equations, and let $\hat{\theta}^a = (\hat{\theta}_1^a, \hat{\theta}_2^a, \hat{\theta}_3^a, \hat{\theta}_4^a)^\top$ denote the solution to $\Psi_J^a(\theta^a) = \sum_{j=1}^J \psi^a(O_j; \theta^a) = 0$. We have that

$$\hat{\theta}_1^a = \frac{1}{\sum_j \sum_{i \in j} R_{ij}} \sum_j \sum_{i \in j} R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij} = \frac{1}{n} \sum_{i=1}^n w_1^a(X_{ij}) Z_{ij} Y_{ij},$$

($i = 1, 2, \dots, n$ represent the enrolled individuals in the J clusters)

$$\hat{\theta}_2^a = \frac{1}{N} \sum_{i=1}^N w_0^a(X_{ij}) (1 - Z_{ij}) Y_{ij}$$

$$\hat{\theta}_3^a = \frac{1}{N} \sum_{i=1}^N w_1^a(X_{ij}) Z_{ij}$$

$$\hat{\theta}_4^a = \frac{1}{N} \sum_{i=1}^N w_0^a(X_{ij}) (1 - Z_{ij}).$$

We similarly define $\hat{\theta}^{a,c}$ and $\hat{\theta}^R$, and $\hat{\theta}$ is the concatenation of the three vectors. Then, the causal estimators can be written as

$$\hat{\tau}_a^O = \frac{\hat{\theta}_1^a}{\hat{\theta}_3^a} - \frac{\hat{\theta}_2^a}{\hat{\theta}_4^a}, \quad \hat{\tau}_{a,c}^O = \frac{\hat{\theta}_1^{a,c}}{\hat{\theta}_3^{a,c}} - \frac{\hat{\theta}_2^{a,c}}{\hat{\theta}_4^{a,c}}, \quad \text{and} \quad \hat{\tau}^R = \frac{\hat{\theta}_1^R}{\hat{\theta}_3^R} - \frac{\hat{\theta}_2^R}{\hat{\theta}_4^R}.$$

Showing that the theorem's conditions hold

1. We show that $\hat{\theta}$ is consistent for θ_0 . For $k, l \in \{1, 2, \dots, 12\}$ we have that

$$\frac{\partial}{\partial \theta_l} \psi_k(O_j; \kappa) = \begin{cases} -\sum_{i \in j} R_i < 0, & \text{if } k = l \\ 0, & \text{otherwise} \end{cases}$$

which implies that θ_0 and $\hat{\theta}$ are both unique roots of the corresponding population and sample level equations (if $\alpha = \alpha_0$). The uniqueness of the roots implies that $\hat{\theta}$ is consistent for θ_0 [Lemma A, Section 7.2.1, Serfling, 2009].

2. Since $\psi(O_j; \theta)$ is linear in θ_k , for all k , we have that $\psi(O_j; \theta)$ is twice continuously differentiable as a function of θ .
3. We want to show that $E \|\psi(O_j; \theta_0)\|_2^2 < \infty$. Take the first entry of the vector:

$$\begin{aligned} E [\psi_1(O_j; \theta_0)^2] &= E \left[\left| \sum_{i \in j} R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij} - \theta_1 \right|^2 \right] \\ &\leq E \left\{ \left[\sum_{i \in j} |R_{ij} w_1^a(X_{ij}) Z_{ij} Y_{ij}| + |\theta_1| \right]^2 \right\} \\ &\leq E \left\{ \left[\sum_{i \in j} w_1^a(X_{ij}) |Y_{ij}| + |\theta_1| \right]^2 \right\} \\ &< c, \end{aligned}$$

for some constant c because $w_1^a(x)$ are bounded since $e(x)$ is bounded from Proposition S.5, and the outcome and cluster size N_j are bounded with probability 1 from Assumption S.1. Similarly we can show that $E [\psi_k(O_j; \theta_0)^2]$ are bounded for all k . Combined, these imply that $E \|\psi(O_j; \theta_0)\|_2^2 < \infty$.

4. From

$$\frac{\partial}{\partial \theta_l} \psi_k(O_j; \kappa) = \begin{cases} -\sum_{i \in j} R_i < 0, & \text{if } k = l \\ 0, & \text{otherwise} \end{cases}$$

for k, l , we have that

$$E \left[\frac{\partial}{\partial \theta^T} \psi(O_j; \theta) \Big|_{\theta_0} \right] = -E[n_j] I_{12},$$

exists and is non-singular since N_j and therefore n_j is bounded from Assumption S.1.

5. Since all second-order derivatives of the estimating function are equal to 0, the integrable function $\ddot{\psi}(o_j) = 0$ dominates all second-order partial derivatives for all θ in a neighborhood of θ_0 .

From Theorem 5.41 of Van der Vaart [2000], we have that

$$\sqrt{J} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) \rightarrow N(0, \Sigma(\boldsymbol{\theta}_0)) \quad \text{as } J \rightarrow \infty,$$

for $\Sigma(\boldsymbol{\theta}_0) = A(\boldsymbol{\theta}_0)^{-1} V(\boldsymbol{\theta}_0) [A(\boldsymbol{\theta}_0)^{-1}]^T$, where

$$A(\boldsymbol{\theta}_0) = E \left[-\frac{\partial}{\partial \boldsymbol{\theta}^T} \psi(O_j; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} \right] = E(-n_j) I_{12}, \quad \text{and} \quad V(\boldsymbol{\theta}_0) = E \left[\psi(O_j; \boldsymbol{\theta}_0) \psi(O_j; \boldsymbol{\theta}_0)^T \right].$$

Consider the multivariate function $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}^a, \boldsymbol{\theta}^{a,c}, \boldsymbol{\theta}^R)$ defined as

$$g(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\theta_1^a}{\theta_3^a} - \frac{\theta_2^a}{\theta_4^a} \\ \zeta_1 \left(\frac{\theta_1^{a,c}}{\theta_3^{a,c}} - \frac{\theta_2^{a,c}}{\theta_4^{a,c}} \right) + \zeta_2 \left(\frac{\theta_1^a}{\theta_3^a} - \frac{\theta_2^a}{\theta_4^a} \right) \\ \frac{\theta_1^R}{\theta_3^R} - \frac{\theta_2^R}{\theta_4^R} \end{pmatrix},$$

for constants $\zeta_1 = [1 - \pi_a/(\pi_a + \pi_c)]^{-1}$ and $\zeta_2 = -\zeta_1(1 - \zeta_1^{-1})$. Then, $g(\hat{\boldsymbol{\theta}}) = (\hat{\tau}_a^O, \hat{\tau}_c^O, \hat{\tau}^R)^\top$, where $\hat{\tau}_c^O$ is calculated using the true proportion of always-recruited among the always- and incentivized-recruited, $\pi_a/(\pi_a + \pi_c)$, and $g(\boldsymbol{\theta}_0) = (\tau_a^O, \tau_c^O, \tau^R)^\top$. Using the delta method,

$$\sqrt{J} \left((\hat{\tau}_a^O, \hat{\tau}_c^O, \hat{\tau}^R)^\top - (\tau_a^O, \tau_c^O, \tau^R)^\top \right) \rightarrow N(0, S),$$

where $S = (\nabla g(\boldsymbol{\theta}_0))^T \Sigma(\boldsymbol{\theta}_0) \nabla g(\boldsymbol{\theta}_0)$ and $\nabla g(\boldsymbol{\theta}_0) =$

$$\begin{pmatrix} \frac{1}{\theta_{03}^a} & -\frac{1}{\theta_{04}^a} & -\frac{\theta_{01}^a}{\theta_{03}^{a,2}} & \frac{\theta_{02}^a}{\theta_{04}^{a,2}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \zeta_2 & -\zeta_2 & -\frac{\zeta_2 \theta_{01}^a}{\theta_{03}^{a,2}} & \frac{\zeta_2 \theta_{02}^a}{\theta_{04}^{a,2}} & \zeta_1 & -\zeta_1 & -\frac{\zeta_1 \theta_{01}^{a,c}}{\theta_{03}^{a,c,2}} & \frac{\zeta_1 \theta_{02}^{a,c}}{\theta_{04}^{a,c,2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\theta_{03}^R} & -\frac{1}{\theta_{04}^R} & -\frac{\theta_{01}^R}{\theta_{03}^{R,2}} & \frac{\theta_{02}^R}{\theta_{04}^{R,2}} \end{pmatrix}.$$

□

Extensions to the asymptotic results First, the asymptotic results of consistency and asymptotic normality can be extended to accommodate the estimated propensity score from a correctly specified parametric model. If α is the vector of parameters, one needs to extend the existing estimating equations in our proof to include the derivative of the log pseudo-likelihood contribution for cluster j , $\psi_\alpha(O_j; \alpha) = \partial \left\{ \sum_{i \in j} R_{ij} \log [e(X_{ij}; \alpha)^{Z_{ij}} (1 - e(X_{ij}; \alpha))^{1-Z_{ij}}] \right\} / \partial \alpha$.

Secondly, the results extend to the case where $\pi_a/(\pi_a + \pi_c)$ is estimated. First, in randomized experiments, the randomization probability π^t is most often known, and the proportion of treated individuals in the recruited population is estimated by its empirical version. Therefore, we estimate

$$\frac{\widehat{\pi_a}}{\pi_a + \pi_c} = \frac{\pi^t}{1 - \pi^t} \frac{1 - \widehat{p}^t}{\widehat{p}^t} = \frac{\pi^t}{1 - \pi^t} \frac{\sum_j \sum_{i \in j} R_{ij}(1 - Z_{ij})}{\sum_j \sum_{i \in j} R_{ij} Z_{ij}}.$$

Therefore, we can acquire consistency and asymptotic normality of our estimators when the proportion of always-recruited in the always- and incentivized-recruited groups is estimated by extending the estimating equation to include terms $\sum_{i \in j} R_{ij}(1 - Z_{ij})$ and $\sum_{i \in j} R_{ij} Z_{ij}$.

B.8 Sensitivity analysis

Proof of Proposition 1. We consider Γ -sized deviations to the enrollment process due to an unmeasured variable expressed as $\Gamma^{-1} \leq \delta(x) / \delta^*(x, u) \leq \Gamma$. Note also, that, using that $e(x) = \delta(x)/\{\delta(x) + r^{-1}\}$, and the definitions of w_{1ij}, w_{1ij}^*

$$\rho_{ij} = \frac{w_{1ij}^*}{w_{1ij}} = \frac{1 - e^*(x, u)}{e^*(x, u)} \frac{e(x)}{1 - e(x)} = \frac{r\delta(x)}{r\delta^*(x, u)} = \frac{\delta(x)}{\delta^*(x, u)} \in [\Gamma^{-1}, \Gamma].$$

Using ρ_{ij} and w_{1ij} , the estimator which uses the true propensity score and we wish to bound can be written as

$$\frac{\sum_{ij} w_{1ij}^* Z_{ij} Y_{ij}}{\sum_{ij} w_{1ij}^* Z_{ij}} = \frac{\sum_{ij} \rho_{ij} w_{1ij} Z_{ij} Y_{ij}}{\sum_{ij} \rho_{ij} w_{1ij} Z_{ij}}.$$

The problem of maximizing (or minimizing) the last quantity under the constraints $\rho_{ij} \in [\Gamma^{-1}, \Gamma]$ can be transformed to a linear programming problem using the Charnes-Cooper transformation [Charnes and Cooper, 1962]. The linear program under this transformation is the one in the main text, where $\lambda_{ij} = \kappa \rho_{ij}$. \square

Bounds for the causal effect estimator on the always- and incentivized-recruited individuals, $\widehat{\tau}_{a,c}^O$, are acquired similarly. For this estimator, the weights of the treated individuals are equal to 1, and we can focus on bounding the part of the estimator that corresponds to the control individuals. Denote the weights for the controls as $w_{0ij}^* = w_0^*(X_{ij}, U_{ij}) = e^*(X_{ij}, U_{ij})/[1 - e^*(X_{ij}, U_{ij})]$. The following proposition provides an algorithm to acquire the bounds of the estimator for violations of the ignorable recruitment assumption up to Γ .

Proposition S.8. *Maximizing (minimizing) $\tau_0^* = \frac{\sum_{ij} w_{0ij}^*(1 - Z_{ij})Y_{ij}}{\sum_{ij} w_{0ij}^*(1 - Z_{ij})}$ under the Γ -violation of the*

ignorable recruitment assumption is equivalent to solving the linear program that maximizes (minimizing) $\sum_{ij} \lambda_{ij} w_{0ij} (1 - Z_{ij}) Y_{ij}$ with respect to λ_{ij} subject to three constraints: (a) $\kappa \Gamma^{-1} \leq \lambda_{ij} \leq \kappa \Gamma$ (b) $\sum_{ij} \lambda_{ij} w_{0ij} (1 - Z_{ij}) = 1$, and (c) $\kappa \geq 0$, where $w_{0ij} = w_0(X_{ij})$ is the weight of unit i under control and working propensity score $e(x)$, and κ is a parameter of the linear program.

Proof. The proof is very similar to that of Proposition 1. \square

These propositions allow us to bound the Hajék estimators $\hat{\tau}_a^O$ and $\hat{\tau}_{a,c}^O$ under Γ -violations of the ignorable recruitment assumption by solving a linear program, which can be achieved computationally fast. On the other hand, the causal estimator for the effect on the incentivized-recruited group is equal to

$$\hat{\tau}_c^O = \frac{\pi_a + \pi_c}{\pi_c} \hat{\tau}_{a,c}^O - \frac{\pi_a}{\pi_c} \hat{\tau}_a^O.$$

The optimization problem for bounding this quantity under Γ -violation of the ignorable recruitment assumption cannot be immediately written as a linear program, and it is likely to require computationally intensive optimization tools. Instead, we use the bounds acquired for $\hat{\tau}_a^O$ and $\hat{\tau}_{a,c}^O$ to also acquire bounds for $\hat{\tau}_c^O$. Specifically, let l_a, u_a , and $l_{a,c}, u_{a,c}$ be the lower and upper bounds for $\hat{\tau}_a^O$ and $\hat{\tau}_{a,c}^O$, respectively. Then, we use $l_c = \frac{\pi_a + \pi_c}{\pi_c} l_{a,c} - \frac{\pi_a}{\pi_c} u_a$ and $u_c = \frac{\pi_a + \pi_c}{\pi_c} u_{a,c} - \frac{\pi_a}{\pi_c} l_a$ as the lower and upper bounds for $\hat{\tau}_c^O$, respectively. If l_c^* and u_c^* denote the theoretical lower and upper bounds for $\hat{\tau}_c^O$, then $[l_c^*, u_c^*] \subseteq [l_c, u_c]$. Therefore, our bounds are valid in that the causal estimator falls within the designed interval under a Γ -violation of the ignorable recruitment assumption. However, this interval might be wider than necessary. The proof is straightforward, hence omitted.

C. Simulations

C.1 Data generative mechanisms

We consider 36 data generative mechanisms that are combinations of the specifications described in Table 1. Here, we provide the details for how the data are generated. We considered three choices for the number of clusters where $J \in \{200, 500, 800\}$. Across all scenarios we specified that each cluster had 100 individuals in its overall population, $N_j = 100$, though (as described below) at least 50% were never-recruited individuals, and therefore n_j was in general smaller than 50.

1. Cluster treatment was generated by choosing m out of J clusters to receive the treatment, where we used $m/J \in \{0.25, 0.5\}$ representing an imbalanced and a balanced treatment assignment, respectively.

2. We generated three individual level covariates V_{ij} and two cluster level covariates V_j^c , with $X_{ij} = (V_{ij}, V_j^c)$. Covariates $V_{1ij}, V_{3ij}, V_{1j}^c$ were generated from independent standard normal distributions truncated to lie between -3 and 3, and V_{2ij}, V_{2j}^c were generated from independent Bernoulli distributions with probability of success 0.5.
3. We considered outcomes with intra-class correlation. Specifically, we generated cluster-level random effects ϵ_j^{icc} from a $N(0, \sigma^2 \rho)$ distribution. Residual variability was generated as ϵ_{ij} from a $N(0, \sigma^2(1 - \rho))$ distribution. Then, potential outcomes were calculated as

$$Y_{ij}(z) = X_{ij}^T \beta_z^Y + \epsilon_{ij} + \epsilon_j^{icc}.$$

4. The potential enrollment under control $R_{ij}(0)$ was generated from a Bernoulli distribution with a logistic link function and linear predictor $X_{ij}^T \beta_0^R$.
5. Under monotonicity, the potential enrollment values under treatment are generated conditional on the potential enrollment values under control. We set $R_{ij}(1) = 1$ for those with $R_{ij}(0) = 1$. We consider parameters α and $\delta(x; \alpha) = 1 + \exp\{-x^t \alpha\}$. Then, for a unit with $R_{ij}(0) = 1$ we generate the corresponding $R_{ij}(1)$ from a Bernoulli distribution with probability of success $[\delta(X_{ij}; \alpha) - 1] \exp\{X_{ij}^T \beta_0^R\}$. Doing so ensures that $\delta(x; \alpha)$ is in fact equal to the ratio in Assumption 3A which is a ratio of marginal probabilities for the recruitment under the two treatments.

We considered data generative mechanisms that led to different prevalence of the principal strata. We named these Scenarios A, B, and C. The proportions of the principal strata are shown on Table 1. We also considered sub-cases that varied how strongly the covariates separate the always-recruited from the group of always- and incentivized-recruited. We named these Cases 1 and 2. We did this by varying the size of the coefficients for the covariates in the model for $\delta(x)$ since

$$\delta(x)^{-1} = P(S = a \mid S \in \{a, c\}, X).$$

Therefore, larger (in absolute value) coefficients for x in $\delta(x)$ lead to stronger covariate separation between the always and the incentivized-recruited.

We found parameters that achieve these specifications by noting that increasing the intercept in $\delta(x; \alpha)$ does not affect the proportion of the population that is always-recruited, whereas it decreases the proportion of incentivized-recruited and increases the proportion of never-recruited. Also, we noted that increasing the intercept in the model for $R(0)$ (the intercept in β_0^R) reduces

Table S.2: Coefficients in the data generative mechanisms in the simulations. These coefficients are chosen to achieve target principal strata distribution in the overall population and vary the covariate separation among always- and incentivized-recruited.

	β_0^R		α
Scenario A	(-0.99, 0.3, -0.6, 0, 0.1, -0.3)	Case 1	(0.275, 0.3, -0.5, -0.1, 0, -0.15)
		Case 2	(0.160, 0.2, -0.3, -0.1, 0, -0.15)
Scenario B	(-0.70, 0.3, -0.6, 0, 0.1, -0.3)	Case 1	(0.275, 0.3, -0.5, -0.1, 0, -0.15)
		Case 2	(0.160, 0.2, -0.3, -0.1, 0, -0.15)
Scenario C	(-1.35, 0.3, -0.6, 0, 0.1, -0.3)	Case 1	(-0.235, 0.3, -0.5, -0.1, 0, -0.15)
		Case 2	(-0.340, 0.2, -0.3, -0.1, 0, -0.15)

the number of never recruited, but does not affect the proportion of always-recruited among the always- and incentivized-recruited.

We found that the coefficients listed in Table S.2 achieve these goals in terms of principal strata distribution. We also set: (a) $\beta_1^Y = (2, -1, 3, 0.1, -0.1, 0.3)$, (b) $\beta_0^Y = (0, -0.5, 1, 0.1, -0.2, 0.3)$, (c) $\sigma^2 = 1$, and (d) $\rho = 0.1$.

C.2 Additional results

Figure S.1 shows the distribution of treatment effects among the different populations of interest (everyone in the overall population, the recruited population, always- and disincentivized-recruited strata, always- and incentivized- recruited strata, and the incentivized-recruited strata) and the distribution for the difference of mean outcomes among treated and control recruited individuals. The latter quantity corresponds to the naïve estimator that ignores recruitment bias. The naïve estimator does *not* return quantities that represent a causal effect over an interpretable population.

Figure S.2 shows the estimated minus the true causal effect on the the three populations of interest, the recruited population, the always-recruited and the incentivized-recruited population, when using the proposed estimators with estimated working propensity score and imbalanced designs that assign 25% of the clusters to treatment. These results show that the estimators are essentially unbiased across all scenarios considered. Similarly, Figure S.3 shows the same quantity across all simulation configurations and when using the known working propensity score. As expected, the estimator of the causal effect in the three populations is unbiased, and it is more precise under a larger number of clusters.

Also, Figure S.4 shows the coverage of the 95% confidence intervals for the causal estimators for the effect on the recruited, the always-recruited, the combination of always- and incentivized-recruited, and the incentivized-recruited populations. For the estimators that use the known

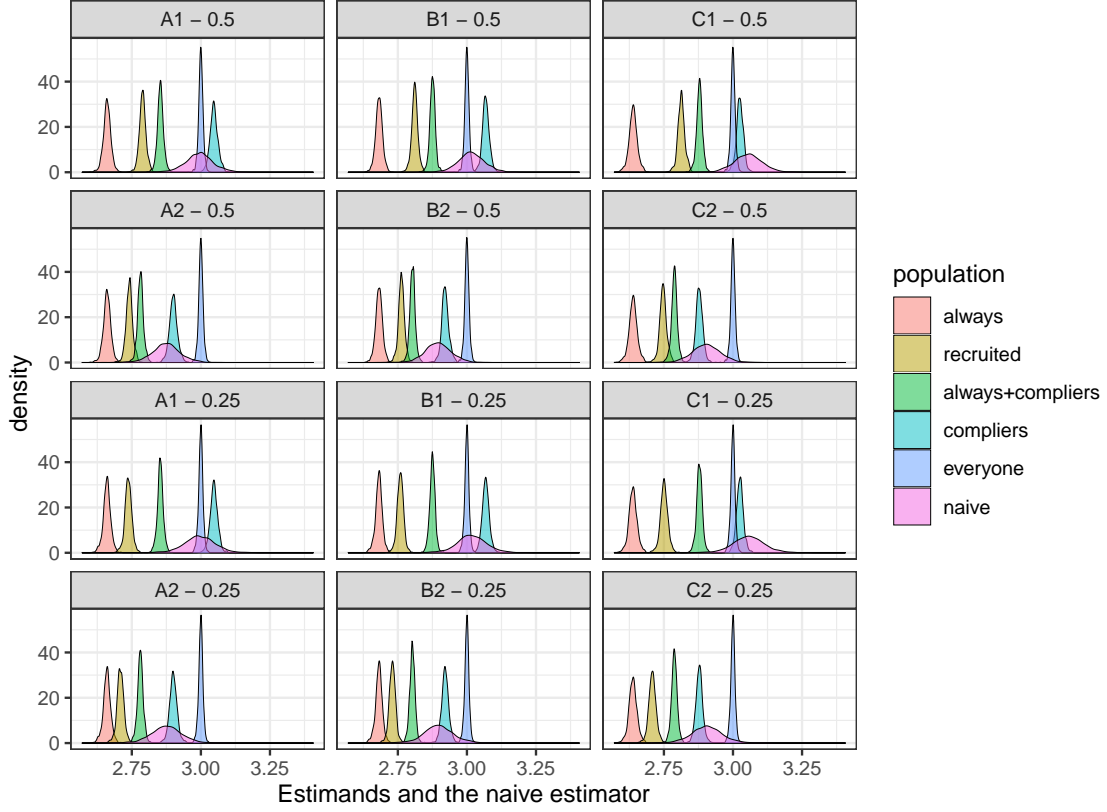


Figure S.1: Distribution of sample causal effect among different populations and estimated naïve difference of mean outcomes across 500 simulated data sets and number of clusters $J \in \{200, 500, 800\}$ organized by the 12 combinations of scenario, case, and treatment proportion in our configurations.

propensity score, we consider 95% intervals based on the derived asymptotic distribution. For the estimators that use the estimated propensity score, we consider a bootstrap procedure that resamples clusters while holding the number of treated clusters fixed, in order to emulate cluster treatment assignment under the simulation design. We calculate the standard deviation of the bootstrap estimates and use them to construct 95% confidence intervals. The coverage of the intervals based on the asymptotic distribution for the estimator that uses the known propensity score is close to 95% across all scenarios. Coverage based on the bootstrap is close to 95% for the estimator of the causal effect on the recruited population, the always-recruited individuals in the overall population, and the combination of always- and incentivized-recruited individuals in the overall population when using the estimated propensity score. Coverage of the same intervals for the causal effect on the incentivized-recruited group are lower, ranging from 85 to 93%, though always closer to the nominal level under Scenario C compared to Scenarios A and B.

Lastly, we investigate the estimation technique of our working propensity score model paramete-

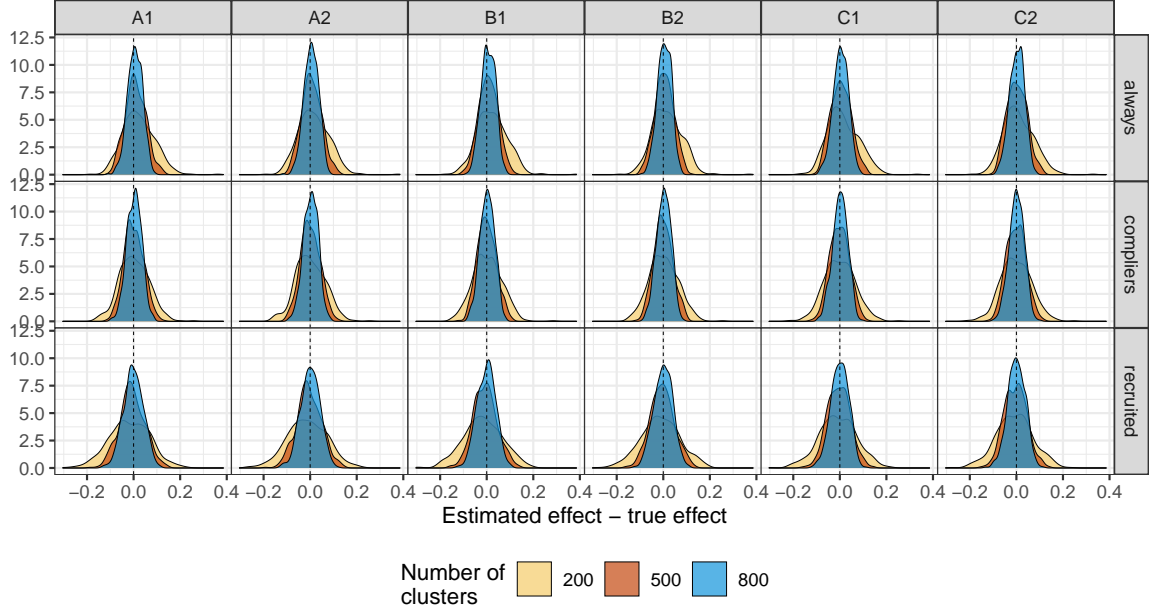


Figure S.2: Bias of the causal estimator for the treatment effect among the always-recruited, the incentivized-recruited, and the recruited populations, across 500 data sets and under 6 scenarios and 3 choices for the number of clusters, when the probability of cluster treatment is 0.25.

ters. In Figure S.5 we show the distribution of estimated minus true parameter across the simulated data sets for scenario A, case 1, and under different number of clusters. Results from the remaining simulation configurations were similar. We see that the working propensity score estimation technique based on the pseudo-likelihood returns unbiased estimates of the working propensity score parameters, that are closer concentrated near the true value for a larger number of clusters.

D. Simulations under violation of Assumption 3B

In Proposition S.1, we established that Assumption 3B is equivalent to $S \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X$. Then, in Proposition S.3, we showed that Assumption 3A and Assumption 5 imply that $S \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X, S \in \{a, c\}$. Since (under monotonicity) the recruited group corresponds to always- and incentivized-recruited individuals only, it is reasonable to wonder whether Assumption 3A and Assumption 5 suffice for identifying overall causal effects.

We performed simulations to investigate this. These simulations mimic the scenarios in the simulations of Section 6 detailed in Supplement C, except they are designed to represent violations of the ignorable recruitment Assumption 3B, while Assumption 3A and Assumption 5 still hold.

Specifically, data are generated under all 36 data generative mechanisms detailed in Supplement

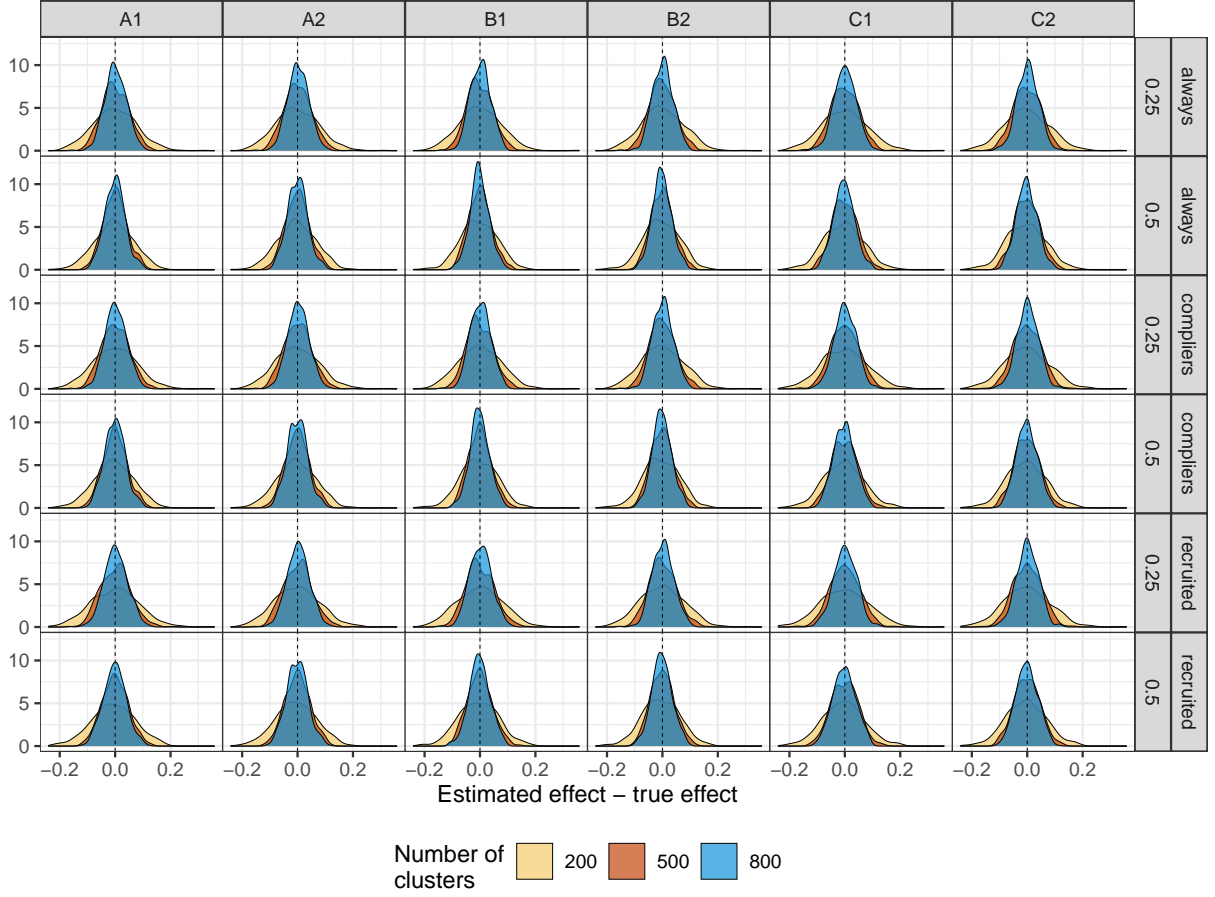


Figure S.3: Bias of the causal estimator that uses the known working propensity score for the treatment effect among the always-recruited, the incentivized-recruited, and the recruited populations, across 500 data sets and under 6 configurations, 3 choices for the number of clusters, and treatment of clusters set to 0.25 or 0.5.

C.1. The main difference lies on the generation of $R_{ij}(0)$ which is now allowed to depend on the values of the potential outcome. Specifically, R_{ij} is generated from a Bernoulli distribution with a logistic link function and linear predictor $(X_{ij}^\top Y_{ij}(0) Y_{ij}(1))^\top \tilde{\beta}_0^R$. Assumption 3A still holds, meaning that the ratio of recruitment under different treatment vectors still only depends on measured covariates X and not potential outcomes $Y(0), Y(1)$, therefore it can still be denoted as $\delta(x)$. We ensured the same prevalence of principal strata and treatment in the recruited group for the simulations where Assumption 3B holds (Section 6) and for the simulations where Assumption 3B does not hold. To do so, we had to slightly alter the intercepts for the models used to generating the recruitment indicators (intercept in $\tilde{\beta}_0^R$ and α) while the coefficients of all measured covariates were kept the same in the two scenarios. The new coefficients are shown in Table S.3. Everything else remained the same.

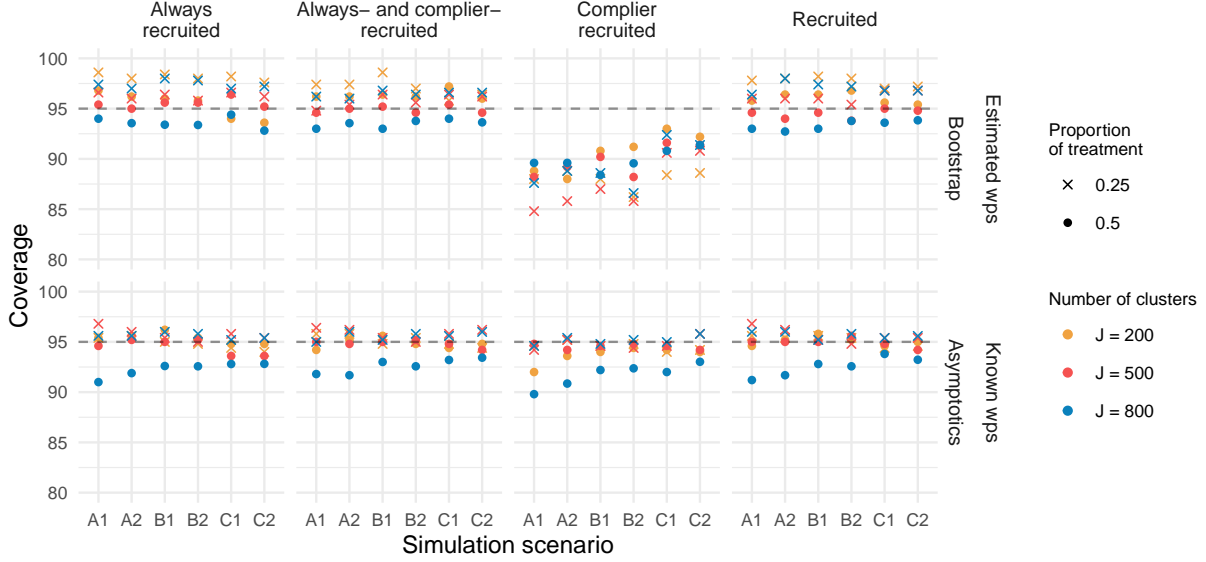


Figure S.4: Coverage of 95% intervals for the estimator of the effect on the recruited, the always-recruited, and the incentivized-recruited populations, based on the asymptotic distribution for the known working propensity score and the bootstrap for estimated propensity score.

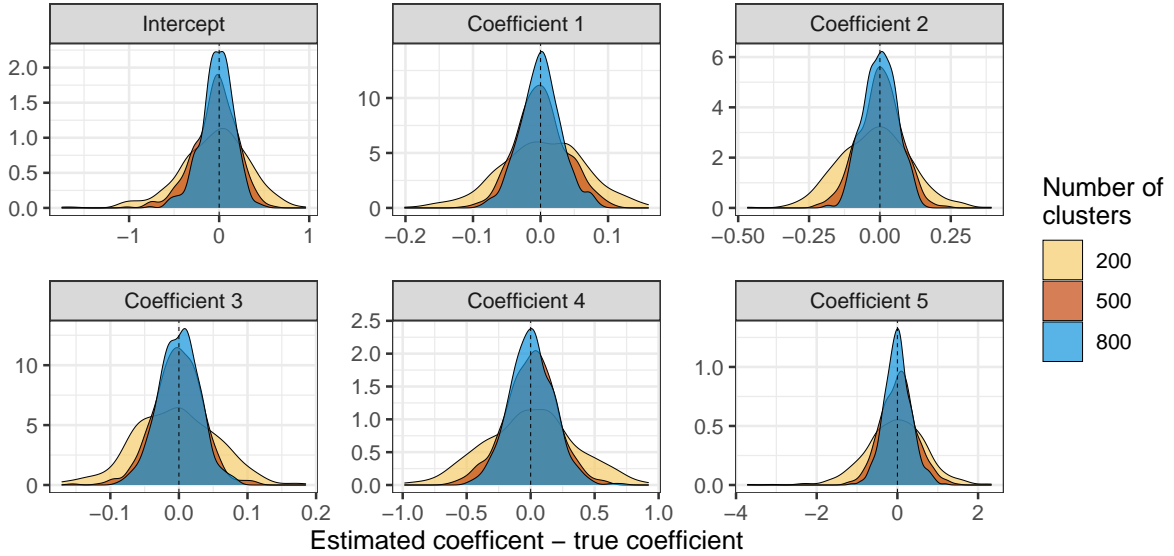


Figure S.5: Distribution of estimated minus true working propensity score coefficients for scenario A, case 1 across simulated data sets, and for different number of clusters.

We show the bias of the proposed estimator based on the estimated working propensity score in Figure S.6. We exclude estimates for the recruited population since identifiability of such effects only requires Assumption 3A that holds. Instead, we focus on the treatment effect on the always-recruited and the incentivized-recruited groups, since these correspond to causal effects over

Table S.3: Coefficients in the data generative mechanisms in the simulations under violations of the ignorable recruitment assumption. These coefficients are chosen to achieve similar data structure to the ones in the simulations of Section 6. The last two entries in $\tilde{\beta}_0^R$ correspond to the potential outcomes.

	$\tilde{\beta}_0^R$		α
Scenario A	(-0.7 , 0.3, -0.6, 0, 0.1, -0.3, 0.2, -0.12)	Case 1	(0.270, 0.3, -0.5, -0.1, 0, -0.15)
		Case 2	(0.160, 0.2, -0.3, -0.1, 0, -0.15)
Scenario B	(-0.4 , 0.3, -0.6, 0, 0.1, -0.3, 0.2, -0.12)	Case 1	(0.272, 0.3, -0.5, -0.1, 0, -0.15)
		Case 2	(0.177, 0.2, -0.3, -0.1, 0, -0.15)
Scenario C	(- 1.07, 0.3, -0.6, 0, 0.1, -0.3, 0.2, -0.12)	Case 1	(-0.268, 0.3, -0.5, -0.1, 0, -0.15)
		Case 2	(-0.355, 0.2, -0.3, -0.1, 0, -0.15)

subgroups of the overall population. We find that our causal estimators are essentially unbiased irrespective of the number of clusters, treatment prevalence, and scenario considered. This provides some indication that the assumptions of the non-differential recruitment (Assumption 3A) and monotonicity (Assumption 5) might suffice for identification of the effects in the always- and incentivized-recruited populations. Even though we considered a large number of simulation scenarios, simulations cannot unequivocally establish that identifiability holds, and further research is required.

E. Information on the ARTEMIS trial

Table S.4 provides descriptive information for the enrolled group of patients with treatment (intervention) and without (control). We perform t-tests for continuous covariates and χ^2 -tests for binary covariates to compare the mean of each covariate in the intervention and control groups. We find that multiple covariates have p-values for the corresponding test that are below 0.05, indicating that the treated and control enrolled groups are different with respect to these characteristics.

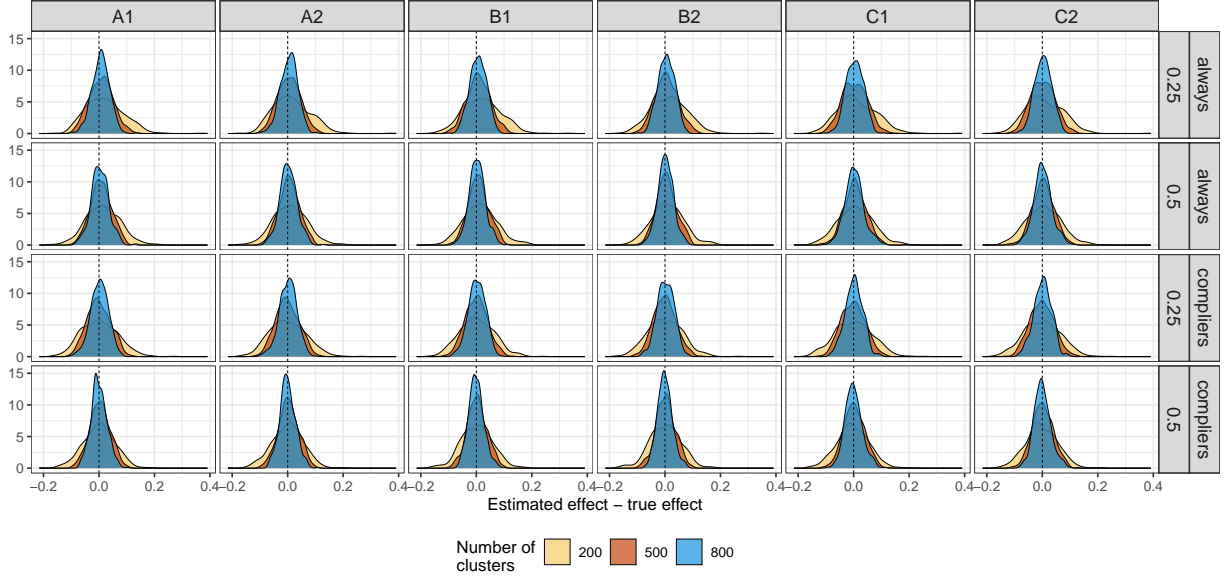


Figure S.6: Bias of the causal estimator that uses the estimated working propensity score under violation of the ignorable recruitment assumption. We show results for the treatment effect among the always-recruited and the incentivized-recruited across 500 data sets and under 6 configurations, 3 choices for the number of clusters, and treatment of clusters set to 0.25 or 0.5.

Covariate	Intervention	Control	<i>p</i> value
Age (year)	62.0 (11.7)	61.4 (11.5)	0.013
Male (%)	68.9 (46.3)	68.3 (46.5)	0.509
White (%)	90.0 (30.1)	86.4 (34.3)	0.000
Black (%)	8.5 (27.8)	11.3 (31.6)	0.000
College (%)	49.4 (50.0)	53.5 (49.9)	0.000
Private insurance (%)	63.2 (48.2)	65.5 (47.5)	0.018
Prior MI (%)	19.3 (39.5)	21.3 (41.0)	0.015
Prior P2Y ₁₂ use (%)	12.6 (33.2)	16.3 (36.9)	0.000
Hemoglobin level	12.87 (2.02)	12.80 (2.08)	0.073
Hypertension (%)	67.2 (47.0)	70.9 (45.4)	0.000
DES use (%)	82.8 (37.8)	78.6 (41.0)	0.000
Diabetes (%)	31.1 (46.3)	34.3 (47.5)	0.000
Employment (%)	48.3 (50.0)	46.9 (49.9)	0.180
Multivessel disease (%)	47.1 (49.9)	45.5 (49.8)	0.133
CABG (%)	1.5 (12.0)	1.4 (11.6)	0.800

Table S.4: Mean Baseline Characteristics of Enrolled Patients by Randomized Arm, with standard deviation in brackets. The *p*-value is computed by *t*-test for continuous data and Pearson's chi-squared test for binary data.

References

- Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. ISSN 1537274X. doi: 10.1080/01621459.1996.10476902.
- Peter M Aronow and Donald KK Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2013.
- Jaime Bolzern, Nandi Mnyama, Katharine Bosanquet, and David J Torgerson. A review of cluster randomized trials found statistical evidence of selection bias. *Journal of clinical epidemiology*, 99:106–112, 2018.
- Gwen Brierley, Sally Brabyn, David Torgerson, and Judith Watson. Bias in recruitment to cluster randomized trials: a review of recent publications. *Journal of evaluation in clinical practice*, 18(4):878–886, 2012.
- Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research logistics quarterly*, 9(3-4):181–186, 1962.
- Peng Ding and Jiannan Lu. Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):757–777, 2017.
- Christina Easter, Jennifer A Thompson, Sandra Eldridge, Monica Taljaard, and Karla Hemming. Cluster randomized trials of individual-level interventions were at high risk of bias. *Journal of clinical epidemiology*, 138:49–59, 2021.
- Constantine E. Frangakis and Donald B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, mar 2002. ISSN 0006341X. doi: 10.1111/j.0006-341X.2002.00021.x.
- Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, 9(1):26–48, 2022.
- Brennan C Kahan, Fan Li, Andrew J Copas, and Michael O Harhay. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology*, 52(1):107–118, 2023.

- C Leyrat, A Caille, A Donner, and B Giraudeau. Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in Medicine*, 32(19):3357–3372, 2013.
- Clémence Leyrat, Agnès Caille, Allan Donner, and Bruno Giraudeau. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Statistics in Medicine*, 33(20):3556–3575, 2014.
- F Li, K L Morgan, and A M Zaslavsky. Balancing covariate via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.
- Fan Li, Zizhong Tian, Jennifer Bobb, Georgia Papadogeorgou, and Fan Li. Clarifying selection bias in cluster randomized trials. *Clinical Trials*, 19(1):33–41, 2022a.
- Fan Li, Zizhong Tian, Zibo Tian, and Fan Li. A note on identification of causal effects in cluster randomized trials with post-randomization selection bias. *Communications in Statistics-Theory and Methods*, pages 1–13, 2022b.
- Roderick Little and Linda Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, pages 1324–1333, 1996.
- Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.
- Georgia Papadogeorgou, Kosuke Imai, Jason Lyall, and Fan Li. Causal inference with spatio-temporal data: estimating the effects of airstrikes on insurgent violence in iraq. *Journal of Royal Statistical Society, Series B*, 2022.
- Suezann Puffer, David Torgerson, and Judith Watson. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *Bmj*, 327(7418):785–789, 2003.
- Paul R Rosenbaum. The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *Journal of the Royal Statistical Society. Series A*, 147(5):656–666, 1984.
- Paul R. Rosenbaum. *Observational studies*. Springer, 2002.

- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- Peter Z Schochet. Estimating complier average causal effects for clustered RCTs when the treatment affects the service population. *Journal of Causal Inference*, 10(1):300–334, 2022.
- Peter Z Schochet. Estimating complier average treatment effects for clustered RCTs with recruitment bias. *Statistics in Medicine*, 2024. doi: 10.1002/sim.9957.
- Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.
- Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- Bingkai Wang, Chan Park, Dylan S. Small, and Fan Li. Model-robust and efficient covariate adjustment for cluster-randomized experiments. *Journal of the American Statistical Association*, 0(0):1–13, 2024.
- Tracy Y Wang, L Kaltenbach, others, Kevin J Anstrom, and Petersen E. Effect of medication co-payment vouchers on p2y12 inhibitor use and major adverse cardiovascular events among patients with myocardial infarction: the artemis randomized clinical trial. *Jama*, 321(1):44–55, 2019.