| Statistica Sinica Preprint No: SS-2025-0476 | |
|---|---|
| **Title** | Conformal Causal Inference for Cluster Randomized Trials: Model-robust Inference Without Asymptotic Approximations |
| **Manuscript ID** | SS-2025-0476 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202025.0476 |
| **Complete List of Authors** | Bingkai Wang,<br>Fan Li and<br>Mengxin Yu |
| **Corresponding Authors** | Bingkai Wang |
| **E-mails** | bingkai.w@gmail.com |
| Notice: Accepted author version. | |

# Conformal causal inference for cluster randomized trials: model-robust inference without asymptotic approximations

Bingkai Wang[1,*], Fan Li[2,3], and Mengxin Yu[4,5]

[1]*Department of Biostatistics, School of Public Health, University of Michigan*

[2]*Department of Biostatistics, Yale School of Public Health*

[3]*Center for Methods in Implementation and Prevention Science, Yale School of Public Health*

[4]*Department of Statistics and Data Science, Washington University in St. Louis*

[5]*School of Public Health, Washington University in St. Louis*

*Abstract:* Traditional statistical inference in cluster randomized trials typically invokes the asymptotic theory that requires the number of clusters to approach infinity. In this article, we propose an alternative conformal causal inference framework for analyzing cluster randomized trials that achieves the target inferential goal in finite samples without the need for asymptotic approximations. Different from traditional inference focusing on estimating the average treatment effect, our conformal causal inference aims to provide prediction intervals for the difference of counterfactual outcomes, thereby providing a new decision-making tool for clusters and individuals in the same target population. We prove that this framework is compatible with arbitrary working outcome models—including

ORCID: Bingkai Wang https://orcid.org/0000-0002-9349-2336, Fan Li https://orcid.org/0000-0001-6183-1893, Mengxin Yu https://orcid.org/0000-0002-6818-4083.

data-adaptive machine learning methods that maximally leverage information from baseline covariates, and enjoys robustness against misspecification of working outcome models. Under our conformal causal inference framework, we develop efficient computation algorithms to construct prediction intervals for treatment effects at both the cluster and individual levels, and further extend to address inferential targets defined based on pre-specified covariate subgroups. Finally, we demonstrate the properties of our methods via simulations and a real data application based on a cluster randomized trial for treating chronic pain.

*Key words and phrases:* Conformal prediction, machine learning, individual-level treatment effect, cluster-level treatment effect, finite-sample coverage.

## 1. Introduction

Cluster randomized trials (CRTs, Murray et al., 1998) are randomized studies in which treatment is assigned to groups rather than individuals. They are increasingly used in pragmatic clinical trials to evaluate interventions under real-world conditions. Compared with individually randomized trials, CRTs can reduce treatment contamination and are well-suited for group-level interventions.

### 1.1 Model-robust inference for CRTs

A growing body of work has focused on model-robust inference for CRTs, seeking alternatives to conventional model-based methods in order to pro-

vide stronger guarantees for causal estimands (e.g., Balzer et al., 2015, 2016; Su and Ding, 2021; Balzer et al., 2023; Wang et al., 2024; Benitez et al., 2023). Model robustness refers to the validity of inference without requiring correct specification of the working model. This is particularly important in CRTs, where outcomes may depend nonlinearly on covariates, exhibit non-normal distributions, and display complex correlation structures. Given these challenges, it is often unrealistic to specify a fully correct model in advance, motivating the need for model-robust inferential tools.

Two major approaches have been explored: model-assisted regression and permutation-based inference. Model-assisted regression relies on asymptotic approximations that require the number of clusters to grow large (Balzer et al., 2015, 2016; Su and Ding, 2021; Wang et al., 2026; Balzer et al., 2023; Wang et al., 2024). In practice, however, CRTs frequently include too few clusters for these asymptotic results to hold. Although finite-sample corrections (e.g., degrees-of-freedom adjustments, Hayes and Moulton, 2017) have been proposed, they are largely developed within model-based frameworks. Permutation tests offer an alternative: they provide exact, model-robust inference under the sharp null (Small et al., 2008; Ding and Keele, 2018) and conservative inference under the weak null (Wu and Ding, 2021). Nonetheless, permutation-based confidence intervals often

require nontrivial computation and may be cumbersome to implement in practice (Rabideau and Wang, 2021).

## 1.2 A concise introduction to conformal prediction

In this article, we further develop conformal prediction, originally conceptualized by Vovk et al. (2005), to perform causal inference in CRTs. This approach achieves finite-sample validity and offers a complementary framework for model-robust inference in CRTs. We first provide a concise introduction to conformal prediction; complete details can be found in Tibshirani et al. (2019); Barber et al. (2021) among others.

Conformal prediction constructs covariate-dependent intervals that contain the outcome with a desired probability (e.g., 90%), regardless of the underlying predictive model. Let $(Y_i, X_i)$, $i = 1, \ldots, n$, be i.i.d. from an unknown distribution $\mathcal{P}$. Conformal prediction produces an interval $\widehat{C}(X)$ satisfying $P\{Y_{\text{new}} \in \widehat{C}(X_{\text{new}})\} \geq 0.9$ for an independent draw $(Y_{\text{new}}, X_{\text{new}}) \sim \mathcal{P}$. The key idea relies on the fact that, for exchangeable variables $V_1, \ldots, V_n, V_{\text{new}}$, their ranks are uniformly distributed, implying that the empirical 0.9-quantile $\widehat{q}_{0.9}$ satisfies $P(V_{\text{new}} \leq \widehat{q}_{0.9}) \geq 0.9$. To avoid using the unknown test point in computing $\widehat{q}_{0.9}$, the empirical distribution is formed by replacing $V_{\text{new}}$ with a point mass at $+\infty$. In prediction settings, we take

$V_i = |Y_i - \widehat{f}(X_i)|$, where $\widehat{f}$ is any fitted model, yielding the interval $\widehat{C}(X) = \{y : |y - \widehat{f}(X)| \leq \widehat{q}_{0.9}\}$, which achieves the desired finite-sample guarantee. A data-splitting step is typically used to separate model fitting from interval construction. Because this validity holds for any choice of $\widehat{f}$ and does not rely on asymptotics, conformal prediction is especially valuable when the number of independent units is limited.

Our work builds upon two lines of research in this area: conformal causal inference (Lei and Candès, 2021; Yang et al., 2024; Qiu et al., 2023; Yin et al., 2024; Alaa et al., 2023; Jin et al., 2023) and conformal prediction for hierarchical data (Dunn et al., 2023; Lee et al., 2023). The former develops conformal prediction intervals for individual treatment effects in non-clustered settings, while the latter provides methods for outcome prediction under hierarchical exchangeability. However, neither line of work has been synthesized to deliver model-robust inference for CRTs, and applying conformal inference to CRTs introduces several methodological challenges. First, the multilevel structure of CRTs gives rise to multiple causal estimands, requiring tailored conformal adaptations for each target parameter. Second, inference on subpopulations can be of substantial interest in CRTs (Wang et al., 2024), yet existing conformal methods do not provide subgroup-valid guarantees for clustered data. Third, the test cluster or

individual may or may not have observed outcomes or randomized treatment assignment, requiring case-specific procedures to optimize precision. Addressing these challenges is central to our proposed framework.

## 1.3 Our contribution

Our principal contribution is to develop a conformal causal inference framework specifically tailored to CRTs. Given any working model, we construct prediction intervals for both cluster-level and individual-level treatment effects and establish their finite-sample, model-robust validity under minimal assumptions. The cluster-level estimand quantifies the effect on a new cluster as a whole, whereas the individual-level estimand reflects the effect on a typical individual within a new cluster; both can be relevant depending on the scientific question (Kahan et al., 2023; Li et al., 2025).

The proposed approach enables treatment-effect bracketing that incorporates individual- and/or cluster-level covariates, providing a practical tool for precision decision-making at multiple levels. We further extend our methods to achieve finite-sample guarantees conditional on covariate subgroups and accommodate different forms of test data. Throughout, we integrate modern machine-learning algorithms as base learners to narrow the prediction intervals for treatment effects without compromising validity.

## 2. Notation, assumptions, and targets of inference

### 2.1 Potential outcomes framework

We consider a CRT with $m$ clusters, and each cluster $i$, $i = 1, \ldots, m$, includes $N_i$ individuals. At the cluster level, we define $A_i$ as a binary treatment indicator ($A_i = 1$ for treatment and $A_i = 0$ for control) and $R_i$ as a vector of cluster-level covariates, such as the geographical location. For each individual $j = 1, \ldots, N_i$ in cluster $i$, we denote $Y_{ij}$ as their observed outcome, $X_{ij}$ as a vector of individual-level baseline covariates. The observed data for each cluster are $O_i = \{(Y_{ij}, A_i, X_{ij}, R_i) : j = 1, \ldots, N_i\}$. We adopt the potential outcomes framework and define $Y_{ij}(a)$ as the potential outcome for individual $j$ of cluster $i$ had cluster $i$ been assigned $A_i = a$ for $a = 0, 1$. We assume causal consistency such that $Y_{ij} = A_i Y_{ij}(1) + (1 - A_i) Y_{ij}(0)$. Denoting the complete data vector for each cluster as $W_i = \{U_{ij} : j = 1, \ldots, N_i\}$ with $U_{ij} = (Y_{ij}(1), Y_{ij}(0), X_{ij}, R_i)$, we make the following structural assumptions on $(W_1, \ldots, W_m)$ and $(A_1, \ldots, A_m)$.

**Assumption 1** (*Random sampling of clusters*)**.** Clusters are independently sampled with $N_i \sim \mathcal{P}^N$ and $W_i | N_i \sim \mathcal{P}^{W|N}$ on $W = \{U_{\bullet,j} : j = 1, \ldots, N\}$ with the subscript '$\bullet, j$' denoting the $j$-th individual in $W$.

**Assumption 2** (*Cluster randomization*)**.** Each $A_i$ is independently drawn

from a Bernoulli distribution $\mathcal{P}^A$ with success probability $\pi \in (0,1)$, and $A_i$ is independent of $W_i$.

**Assumption 3** (*Within-cluster exchangeability*)**.** Given $N$, the vectors $U_{\bullet,j} = (Y_{\bullet,j}(1), Y_{\bullet,j}(0), X_{\bullet,j}, R)$ are exchangeable within each cluster; that is, for any permutation $\sigma$ of $\{1,\ldots,N\}$, $\{U_{\bullet,1},\ldots,U_{\bullet,N}\}$ has the same distribution as $\{U_{\bullet,\sigma(1)},\ldots,U_{\bullet,\sigma(N)}\}$.

Assumption 1 is standard for causal inference in CRTs under a sampling-based framework (Wang et al., 2024). Assumption 2 holds by design. Assumption 3 states that, given the cluster size, the joint distribution of the complete data vector in a cluster is invariant to permutation of index $j$. This stronger assumption is only required for inferring the individual-level treatment effect (Section 4) but not the cluster-level treatment effect (Section 3). For additional illustration, we provide in Supplementary Material A an example class of data-generating processes that imply within-cluster exchangeability. It is clear from that example that the conditional correlation of $Y_{\bullet,j}$ and $Y_{\bullet,j'}$ is allowed to flexibly depend on $X_{\bullet,j}$ and $X_{\bullet,j'}$ through a shared function across $(j,j')$ pairs. However, Assumption 3 is still stronger than those invoked in Balzer et al. (2016); Benitez et al. (2023); Wang et al. (2024), which allow for arbitrary intracluster correlation structures. Furthermore, Assumptions 1–3 induce the same marginal hierarchical structure

as that in Lee et al. (2023), although the latter introduces an additional latent layer to accommodate conditional heterogeneity.

## 2.2 Conventional targets of inference in CRTs

Statistical inference for CRTs traditionally targets average treatment effects defined by summaries of the potential outcomes (Su and Ding, 2021; Benitez et al., 2023). A common estimand is the cluster-average treatment effect, $\Delta_C = E\{\overline{Y}(1) - \overline{Y}(0)\}$, where $\overline{Y}(a) = N^{-1}\sum_{j=1}^{N} Y_{\bullet,j}(a)$ denotes the within-cluster average potential outcome. The inferential objective is to construct a confidence interval $\widehat{\mathrm{CI}}_m$ satisfying $P\{\Delta_C \in \widehat{\mathrm{CI}}_m\} = 1-\alpha$, ideally leveraging covariates to improve efficiency.

This objective is difficult to achieve without strong parametric assumptions. Recent work has therefore developed asymptotic theory ensuring $\lim_{m\to\infty} P\{\Delta_C \in \widehat{\mathrm{CI}}_m\} = 1-\alpha$ even under working-model misspecification. However, because CRTs often include only a small number of clusters, such asymptotic guarantees may be inadequate in practice.

## 2.3 Targets of conformal inference in CRTs

To obtain finite-sample, model-robust inference, we introduce alternative inferential targets. For any $\alpha \in (0, 1)$, our goal is to construct a covariate-

based interval $\widehat{\mathrm{C}}_C(\overline{B}) \subset \mathbb{R}$ for the cluster-level treatment effect such that

$$P\{\overline{Y}(1) - \overline{Y}(0) \in \widehat{\mathrm{C}}_C(\overline{B})\} \geq 1 - \alpha, \tag{2.1}$$

where $\overline{B} = (\overline{X}, R, N)$ denotes the cluster-level covariates and $\overline{X}$ is the within-cluster average of individual-level covariates. Under (2.1), a new cluster drawn from $\mathcal{P}^N \times \mathcal{P}^{W|N}$ has its treatment effect covered with probability at least $1 - \alpha$, without relying on asymptotic approximations. This new cluster can be either an existing cluster in the study or a new cluster outside the CRT, provided it is independently sampled from $\mathcal{P}^N \times \mathcal{P}^{W|N}$. Although motivated by cluster averages, the same formulation applies to other cluster summaries, such as totals.

We also target an individual-level conformal interval $\widehat{\mathrm{C}}_I(B)$ satisfying

$$P\{Y(1) - Y(0) \in \widehat{\mathrm{C}}_I(B)\} \geq 1 - \alpha, \tag{2.2}$$

where $B = (X, R, N)$ denotes the individual-level covariates. Since individuals are identically distributed given $N$ under Assumption 3, the interpretation mirrors that of the cluster-level case: for any individual in a new cluster sampled from $\mathcal{P}^N \times \mathcal{P}^{W|N}$, the treatment effect $Y(1) - Y(0)$ lies in $\widehat{\mathrm{C}}_I(B)$ with probability at least $1 - \alpha$.

Beyond marginal guarantees in (2.1)–(2.2), we also study subgroup-valid inference. Let $\Omega_C$ and $\Omega_I$ be subsets of the supports of $\overline{B}$ and $B$, e.g., age-based subgroups such as $\{\overline{X}_1 \geq 70\}$ or $\{X_1 \geq 70\}$. We construct intervals $\widetilde{\mathrm{C}}_C(\overline{B})$ and $\widetilde{\mathrm{C}}_I(B)$ satisfying

$$P\{\overline{Y}(1) - \overline{Y}(0) \in \widetilde{\mathrm{C}}_C(\overline{B}) \mid \overline{B} \in \Omega_C\} \geq 1 - \alpha, \tag{2.3}$$

$$P\{Y(1) - Y(0) \in \widetilde{\mathrm{C}}_I(B) \mid B \in \Omega_I\} \geq 1 - \alpha. \tag{2.4}$$

These "local coverage" targets provide subgroup-specific validity and finer inferential resolution, unlike marginal coverage over the full population. Marginal coverage is recovered by taking $\Omega_C$ and $\Omega_I$ as their full supports.

Finally, we introduce additional notation: $\overline{Y} = A\overline{Y}(1) + (1 - A)\overline{Y}(0)$; $I\{G\}$ denotes the indicator of event $G$; $\|\cdot\|_2$ is the $\ell_2$ norm; for two sets $C_1$ and $C_2$, we define $C_1 - C_2 = \{c_1 - c_2 : c_1 \in C_1, c_2 \in C_2\}$; and $\delta_s$ and $\delta_{+\infty}$ denote point masses at $s$ and $+\infty$, respectively.

## 3. Conformal causal inference for cluster-level treatment effects

### 3.1 Inference for an observed cluster

For inferring cluster-level treatment effects, we aim to construct a conformal interval $\widetilde{\mathrm{C}}_C$ given a test point $\overline{B}_{\text{test}}$, i.e., the cluster-aggregate covari-

ates of a new cluster sampled from the target population. We first consider a basic scenario, where we observe the complete information $\overline{O}_{\text{test}} = (\overline{Y}_{\text{test}}, A_{\text{test}}, \overline{B}_{\text{test}})$ for the test point. In other words, this test point corresponds to an "observed cluster", i.e., its current intervention $A_{\text{test}}$ (likely not randomized, often $A_{\text{test}} = 0$ and not included in the CRT) and current average outcome $\overline{Y}(A_{\text{test}})$ are also recorded. Such additional data beyond $\overline{B}_{\text{test}}$ will simplify inference since one of the two potential outcomes $\{\overline{Y}_{\text{test}}(1), \overline{Y}_{\text{test}}(0)\}$ is directly observed. Since treatment may not be randomized outside the CRT, $A_{\text{test}}$ may not follow $\mathcal{P}^A$ as in Assumption 2. We therefore assume that $\overline{O}_{\text{test}}$ is the cluster average of a new cluster independently sampled from $\mathcal{P}^N \times \mathcal{P}^{W|N} \times \widetilde{\mathcal{P}}^{A|W,N}$, where $\widetilde{\mathcal{P}}^{A|W,N}$ is an arbitrary treatment assignment distribution. When $\widetilde{\mathcal{P}}^{A|W,N} = \mathcal{P}^A$, the prediction target corresponds to clusters within the CRT. In this case, the target cluster should be excluded from the CRT dataset to avoid in-sample prediction.

Under this basic scenario, Algorithm 1 outlines the steps to compute the conformal interval $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}})$. In the input phase, the prediction model $f_a, a \in \{0, 1\}$ can be an arbitrary map from covariates to the outcome, e.g., a linear model, or random forest (Breiman, 2001). Different choices of $f_a$ will not impact the coverage validity, but can result in conformal intervals with different lengths and thus affect precision. In addition, we need to

specify the covariate subgroup of interest, $\Omega_C$, and a level $\alpha$, e.g., $\alpha = 0.1$.

---

**Algorithm 1** Computing the conformal interval $\widetilde{\mathrm{C}}_C(\overline{O})$ for cluster-level treatment effects.

---

**Input:** Cluster-level data $\{(\overline{Y}_i, A_i, \overline{B}_i) : i = 1, \ldots, m\}$, a test point $\overline{O}_{\text{test}} = (\overline{Y}_{\text{test}}, A_{\text{test}}, \overline{B}_{\text{test}})$, an arbitrary prediction model $f_a(\overline{B})$ for $\overline{Y}(a)$, $a \in \{0, 1\}$, a covariate subgroup of interest $\Omega_C$, and level $\alpha$.

**Step 1** Constructing the conformal interval $\widetilde{\mathrm{C}}_{C,a}(\overline{B})$ for $\overline{Y}(a)$.
For $a = 0, 1$,

1. Randomly split the arm-$a$ covariate subgroup data $\{(\overline{Y}_i, \overline{B}_i) : i = 1, \ldots, m, A_i = a, \overline{B}_i \in \Omega_C\}$ into a training fold $\mathcal{O}_{tr}(a)$ and a calibration fold $\mathcal{O}_{ca}(a)$ with index set $\mathcal{I}_{ca}(a)$.
2. Train the prediction model $f_a(\overline{B})$ using the training fold $\mathcal{O}_{tr}(a)$, and obtain the estimated model $\widehat{f}_a(\overline{B})$.
3. For each $i \in \mathcal{I}_{ca}(a)$, compute the non-conformity score $s(\overline{B}_i, \overline{Y}_i) = |\overline{Y}_i - \widehat{f}_a(\overline{B}_i)|$.
4. Compute the $1 - \alpha$ quantile $\widehat{q}_{1-\alpha}(a)$ of the distribution

$$\widehat{F} = \frac{1}{|\mathcal{I}_{ca}(a)| + 1} \left\{ \sum_{i \in \mathcal{I}_{ca}(a)} \delta_{s(\overline{B}_i, \overline{Y}_i)} + \delta_{+\infty} \right\}.$$

5. Obtain $\widetilde{\mathrm{C}}_{C,a}(\overline{B}) = \{y \in \mathbb{R} : |y - \widehat{f}_a(\overline{B})| \leq \widehat{q}_{1-\alpha}(a)\}$.

**Step 2** Constructing the conformal interval $\widetilde{\mathrm{C}}_C(\overline{O})$ for $\overline{Y}(1) - \overline{Y}(0)$.
If $A_{\text{test}} = 1$, then set $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}) = \overline{Y}_{\text{test}} - \widetilde{\mathrm{C}}_{C,0}(\overline{B}_{\text{test}})$;
if $A_{\text{test}} = 0$, then set $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}) = \widetilde{\mathrm{C}}_{C,1}(\overline{B}_{\text{test}}) - \overline{Y}_{\text{test}}$.

**Output:** $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}})$.

---

Given the input, we take two steps to construct the conformal interval. In the first step, we apply the split conformal prediction (Papadopoulos et al., 2002) to construct a conformal interval for one potential outcome $\overline{Y}(a)$, based on arm-$a$ data in the covariate subgroup characterized by $\Omega_C$.

In split conformal prediction, the data are randomly partitioned into two parts, one used to train the prediction model $f_a$, $a \in \{0, 1\}$, (Steps 1.1-1.2) and the other used to construct the conformal interval (Steps 1.3-1.5). Given the model fit $\widehat{f}_a$, Step 1.3 computes the non-conformity score, which is the absolute value of prediction error on calibration data. Intuitively, a large non-conformity score indicates an abnormal data point (i.e., it does not "conform") with respect to $\widehat{f}_a$. In Step 1.4, we construct an empirical distribution of the non-conformity score, denoted as $\widehat{F}$, among the validation samples and the test sample; since the test sample may not be in group $a$ and hence unobserved, we replace its point mass by a point mass at infinity. Following the same idea of conformal prediction described in Section 1.2, we have $P\{s(\overline{B}_{\text{test}}, \overline{Y}_{\text{test}}(a)) \leq \widehat{q}_{1-\alpha}(a)\} \geq 1 - \alpha$, where $\widehat{q}_{1-\alpha}(a)$ is the $(1-\alpha)$-quantile of $\widehat{F}$. This leads to our conformal interval for $\overline{Y}_{\text{test}}(a)$ in Step 1.5. We then proceed to Step 2 and output $\widetilde{\text{C}}_C(\overline{O}_{\text{test}}) = (-1)^{A_{\text{test}}+1}\{\overline{Y}_{\text{test}} - \widetilde{\text{C}}_{C,1-A_{\text{test}}}(\overline{B}_{\text{test}})\}$; that is, the final conformal interval is a contrast between the observed potential outcome and the interval constructed for the unobserved potential outcome.

Theorem 1 proves that $\widetilde{\text{C}}_C(\overline{O}_{\text{test}})$ achieves finite-sample coverage guarantee for the cluster-level treatment effect, without requiring Assumption 3 on within-cluster exchangeability.

**Theorem 1.** *Under Assumptions 1-2, the conformal interval* $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}})$ *output by Algorithm 1 satisfies*

$$P\left\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \in \widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}) \middle| \overline{B}_{\text{test}} \in \Omega_C \right\} \geq 1 - 2\alpha \tag{3.5}$$

*for any set* $\Omega_C$ *in the support of* $\overline{B}_{\text{test}}$ *with a positive measure. If* $\widetilde{\mathcal{P}}^{A|W,N} = \mathcal{P}^A$*, i.e., the test cluster receives randomized treatment, then the coverage rate in Equation* (3.5) *improves to* $1 - \alpha$*.*

Theorem 1 is a direct generalization of the conformal prediction theory to conformal causal inference on covariate subgroups. Since we observe $\overline{Y}_{\text{test}}$ and $A_{\text{test}}$, the valid coverage for the treatment effect is straightforward once we achieve valid coverage for the potential outcome. However, since we need to account for arbitrary distribution shift on $A_{\text{test}}$ from $\mathcal{P}^A$ to $\widetilde{\mathcal{P}}^{A|W,N}$, the non-coverage rate $\alpha$ is doubled in Equation (3.5) by applying the union bound. For local coverage, we consider a simple yet effective approach that only clusters within $\overline{B}_i \in \Omega_C$ are included for analysis (Vovk, 2012), resembling the conventional idea for subgroup analysis. This can change the covariate distribution but not the outcome distribution given covariates, thereby not affecting the coverage result. Beyond this approach, an alternative method that achieves uniform local coverage using the full

sample is discussed in Hore and Barber (2025).

In Algorithm 1, both the test point and the prediction model are based on cluster-level averages. When individual-level data are available for the test cluster, one may instead fit an individual-level prediction model $f_a^*(B_{ij})$ for $Y_{ij}(a)$ and define the non-conformity score as $|\overline{Y}_i - \frac{1}{N_i}\sum_{j=1}^{N_i} \widehat{f}_a^*(B_{ij})|$. This approach leverages all observed covariates, potentially increasing the effective sample size for model training and improving the stability and precision of the nonconformity scores (see Supplementary Material D for a numerical illustration). In practice, the choice between cluster-level summaries and individual-level predictors depends on data availability, domain knowledge, and the complexity of the prediction task.

In practice, how to choose $\alpha$ and the size of the training fold depends on the number of clusters $m$. To obtain a non-trivial conformal interval such that $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}) \neq \mathbb{R}$, the definition of $\widehat{q}_{1-\alpha}(a)$ requires that $\alpha \geq |\mathcal{I}_{ca}(a)|^{-1}$. For example, if we set $\alpha = 0.1$, then the calibration fold for each arm should contain at least 10 clusters, and the rest $m-20$ clusters can be used as the training fold. For stability of model fitting, we recommend using at least 20 clusters for training, while fewer clusters are also acceptable if parsimonious models such as linear regression are used. When the number of clusters is small, e.g., $m = 20$, Algorithm 1 cannot construct non-trivial

conformal intervals with 90% coverage probability. However, this goal may be achievable with full conformal prediction or Jackknife+, which demand substantially heavier computation (Barber et al., 2021).

### 3.2 Inference based on cluster-level covariates

When the target of inference concerns a new cluster that only has covariate information $\overline{B}_{\text{test}}$ (e.g., a new cluster having not taken either treatment or control studied in the current CRT), a direct approach based on Algorithm 1 is to combine $\widetilde{\text{C}}_{C,1}$ and $\widetilde{\text{C}}_{C,0}$, for which Corollary 1 characterizes its local coverage property.

**Corollary 1.** *Under Assumptions 1-2 and assuming that* $(\overline{Y}_{\text{test}}(1), \overline{Y}_{\text{test}}(0), \overline{B}_{\text{test}})$ *is an independent sample from the distribution* $\mathcal{P}^N \times \mathcal{P}^{W|N}$. *Then,* $\widetilde{\text{C}}_{C,1}(\overline{B}_{\text{test}})$ *and* $\widetilde{\text{C}}_{C,0}(\overline{B}_{\text{test}})$ *output by Algorithm 1 satisfy*

$$P\left\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \in \widetilde{\text{C}}_{C,1}(\overline{B}_{\text{test}}) - \widetilde{\text{C}}_{C,0}(\overline{B}_{\text{test}}) \middle| \overline{B}_{\text{test}} \in \Omega_C\right\} \geq 1 - 2\alpha \tag{3.6}$$

*for any set* $\Omega_C$ *in the support of* $\overline{B}_{\text{test}}$ *with a positive measure.*

Compared to Equation (3.5), Equation (3.6) provides the same coverage probability $1 - 2\alpha$, but the length of the conformal interval can be approximately doubled since we no longer observe $\overline{Y}_{\text{test}}$ and $A_{\text{test}}$. As a

result, the conformal interval based on $\overline{B}_{\text{test}}$ tends to be less informative than $\widetilde{\text{C}}_C(\overline{O}_{\text{test}})$, a natural result of less observed information.

Beyond this direct approach, Lei and Candès (2021) provided a more flexible nested approach to construct $\widetilde{\text{C}}_C(\overline{B}_{\text{test}})$. This method first computes $\overline{C}_i$, the $(1-\alpha)$-conformal interval for $\overline{Y}(1)-\overline{Y}(0)$, using $\widetilde{\text{C}}_{C,a}$ from Algorithm 1, and then run split conformal prediction again on $(\overline{C}_i, \overline{B}_i)$ with level $\gamma$ and prediction models $(m^L, m^R)$. Since we run split conformal prediction twice, the resulting conformal interval $\widetilde{\text{C}}_C(\overline{B}_{\text{test}})$ has non-coverage probability up to $\alpha+\gamma$. This approach becomes similar to our direct approach by setting $\gamma=\alpha$. For completeness, we provide the detailed algorithm and its theoretical guarantee in Supplementary Material C.

## 4. Conformal causal inference for individual-level treatment effects

### 4.1 Inference for an observed individual

We first consider a basic setting, where the test point has the complete information $O_{\text{test}} = (Y_{\text{test}}, A_{\text{test}}, B_{\text{test}})$. This setting corresponds to an observed individual in an observed cluster of interest, whose current treatment $A_{\text{test}}$ may not be randomized. Similar to our development in Section 3.1, we assume that $O_{\text{test}}$ is an arbitrary individual from a new cluster indepen-

dently sampled from $\mathcal{P}^N \times \mathcal{P}^{W|N} \times \widetilde{\mathcal{P}}^{A|W,N}$, where $\widetilde{\mathcal{P}}^{A|W,N}$ is an arbitrary unknown distribution for $A_{\text{test}}$. The distribution of interest is therefore the same hierarchical sampling mechanism that generates $W_1, \ldots, W_m$ in the CRTs.

In Algorithm 2, we show how to construct the conformal interval $\widetilde{\text{C}}_I(O)$ for $Y(1) - Y(0)$. Compared to Algorithm 1, the major difference is in Step 1.4: the new empirical distribution function $\widehat{F}$ only involves individuals meeting the criteria $B_{ij} \in \Omega_I$, and each individual is further weighted by $\left(\sum_{j=1}^{N_i} I\{B_{ij} \in \Omega_I\}\right)^{-1}$; all other steps remain similar but now apply to the individual data (rather than cluster aggregate). In this new $\widehat{F}$, we first construct an empirical distribution within each cluster and then construct the empirical distribution connecting validation clusters and the test cluster. This new $\widehat{F}$ captures two levels of exchangeability (within and across clusters), which extends conformal prediction of single-level data (Section 1.2 and Algorithm 1). With this change, we can establish that $P\{s(B_{\text{test}}, Y_{\text{test}}(a)) \leq \widehat{q}_{1-\alpha}(a)\} \geq 1 - \alpha$ and obtain the desired conformal conformal intervals for $Y_{\text{test}}(a)$. Overall, Algorithm 2 generalizes hierarchical conformal prediction (Lee et al., 2023) to target treatment effects (rather than outcomes) and to accommodate covariate subgroup analysis. For the latter, we develop new technical arguments (Lemma 1 in Supplementary

Material B) to show that the hierarchical exchangeability conditions required for validity continue to hold when restricting to covariate-defined subgroups in CRTs. The theoretical guarantee of the resulting conformal interval is formally stated in Theorem 2.

**Theorem 2.** *Under Assumptions 1-3, the* $\widetilde{\mathrm{C}}_I(O_{\text{test}})$ *output by Algorithm 2 satisfies*

$$P\left\{Y_{\text{test}}(1) - Y_{\text{test}}(0) \in \widetilde{\mathrm{C}}_I(O_{\text{test}}) \middle| B_{\text{test}} \in \Omega_I\right\} \geq 1 - 2\alpha \tag{4.7}$$

*for any set* $\Omega_I$ *in the support of* $B_{\text{test}}$ *with a positive measure. If* $\widetilde{\mathcal{P}}^{A|W,N} = \mathcal{P}^A$*, then the coverage rate in Equation* (4.7) *improves to* $1 - \alpha$*.*

Theorem 2 is the counterpart of Theorem 1 for individual-level treatment effects. Due to the distribution shift on $A_{\text{test}}$, the coverage probability is $1-2\alpha$ instead of $1-\alpha$, while the latter can be achieved if $\widetilde{\mathcal{P}}^{A|W,N}$ is independent of $W$ (e.g., $A_{\text{test}} \equiv 0$). When each cluster only has one individual, Algorithm 1 and Algorithm 2 coincide, and their resulting coverage guarantees also become identical.

In terms of the length of conformal intervals, $\widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}})$ tends to be more informative than $\widetilde{\mathrm{C}}_I(B_{\text{test}})$ given a sufficient number of clusters. This is because $\overline{Y}(1) - \overline{Y}(0)$ often has smaller variance than $Y(1) - Y(0)$, espe-

**Algorithm 2** Computing the conformal interval $\widetilde{\mathrm{C}}_I(O)$ for individual-level treatment effects.

**Input:** Individual-level data $\{(Y_{ij}, A_i, B_{ij}) : i = 1, \ldots, m; j = 1, \ldots, N_i\}$, a test point $O_{\text{test}} = (Y_{\text{test}}, A_{\text{test}}, B_{\text{test}})$, a prediction model $f_a(B)$ for $Y(a)$, $a \in \{0, 1\}$, a covariate subgroup of interest $\Omega_I$, and level $\alpha$.

**Step 1** (Constructing the conformal interval $\widetilde{\mathrm{C}}_{I,a}(B)$ for $Y(a)$.)
For $a = 0, 1$,

1. Randomly split the arm-$a$ covariate subgroup data $\{(Y_{ij}, A_i, B_{ij}) : i = 1, \ldots, m; j = 1, \ldots, N_i; A_i = a; B_{ij} \in \Omega_I\}$ into a training fold $\mathcal{O}_{tr}(a)$ and a calibration fold $\mathcal{O}_{ca}(a)$ with index set $\mathcal{I}_{ca}(a)$. The split is at the cluster level, and individuals in the same cluster remain in the same fold.
2. Train the prediction model $f_a(B)$ using the training fold $\mathcal{O}_{tr}(a)$, and obtain the estimated model $\widehat{f}_a(B)$.
3. For each $(i, j) \in \mathcal{I}_{ca}(a)$, compute the non-conformity score $s(B_{ij}, Y_{ij}) = |Y_{ij} - \widehat{f}_a(B_{ij})|$.
4. Compute the $1 - \alpha$ quantile $\widehat{q}_{1-\alpha}(a)$ of the distribution

$$\widehat{F} = \frac{1}{|\mathcal{I}_{ca}(a)| + 1} \left\{ \sum_{i \in \mathcal{I}_{ca}(a)} \frac{1}{\sum_{j=1}^{N_i} I\{B_{ij} \in \Omega_I\}} \sum_{j=1}^{N_i} I\{B_{ij} \in \Omega_I\} \delta_{s(B_{ij}, Y_{ij})} + \delta_{+\infty} \right\}.$$

5. Obtain $\widetilde{\mathrm{C}}_{I,a}(B) = \{y \in \mathbb{R} : |y - \widehat{f}_a(B)| \leq \widehat{q}_{1-\alpha}(a)\}$.

**Step 2** (Constructing the conformal interval $\widetilde{\mathrm{C}}_I(O)$ for $Y(1) - Y(0)$.)
If $A_{\text{test}} = 1$, then set $\widetilde{\mathrm{C}}_I(O_{\text{test}}) = Y_{\text{test}} - \widetilde{\mathrm{C}}_{I,0}(B_{\text{test}})$;
if $A_{\text{test}} = 0$, then set $\widetilde{\mathrm{C}}_I(O_{\text{test}}) = \widetilde{\mathrm{C}}_{I,1}(B_{\text{test}}) - Y_{\text{test}}$.

**Output:** $\widetilde{\mathrm{C}}_I(O_{\text{test}})$.

cially if $N_i$ is large. As a result, $\widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}})$ is more likely to exclude zero than $\widetilde{\mathrm{C}}_I(B_{\text{test}})$ given the same treatment effect size. In practice, choosing between the two types of inferential targets requires a case-by-case evaluation. Although the scientific question should drive the target of inference, from a statistical perspective, conformal inference for the cluster-level treatment effects is typically more informative when $m$ is large, e.g., $m \geq 80$. Given a small to moderate number of clusters, conformal causal inference for individual-level treatment effects could be numerically more stable due to the increased sample size in the calibration fold to compute $\widehat{q}_{1-\alpha}(a)$.

## 4.2 Inference based on individual-level covariates

When the test point only contains individual-level covariates $B_{\text{test}}$, we follow the same strategy as in Section 3.2 to construct conformal intervals. Corollary 2 characterizes the local coverage property of the direct approach as an application of Theorem 2.

**Corollary 2.** *Under Assumptions 1-3 and assuming that* $(Y_{\text{test}}(1), Y_{\text{test}}(0), B_{\text{test}})$ *is an arbitrary individual from a new cluster independently sampled from* $\mathcal{P}^N \times \mathcal{P}^{W|N}$. *Then* $\widetilde{\mathrm{C}}_{I,1}(B_{\text{test}})$ *and* $\widetilde{\mathrm{C}}_{I,0}(B_{\text{test}})$ *output by Algorithm 2 satisfy*

$$P\left\{Y_{\text{test}}(1) - Y_{\text{test}}(0) \in \widetilde{\mathrm{C}}_{I,1}(B_{\text{test}}) - \widetilde{\mathrm{C}}_{I,0}(B_{\text{test}}) \middle| B_{\text{test}} \in \Omega_I\right\} \geq 1 - 2\alpha \quad (4.8)$$

*for any set $\Omega_I$ in the support of $B_{\text{test}}$ with a positive measure.*

In parallel to Corollary 1, Corollary 2 establishes the coverage guarantee on conformal intervals for the individual-level treatment effect based on covariates. In addition, the resulting conformal interval enjoys the same benefit of stability and small-sample compatibility as discussed in Section 4.1. In addition to the direct approach, we provided the nested approach with $1-\alpha-\gamma$ coverage guarantee in Supplementary Material C.

## 5. Simulations

### 5.1 Simulation design

Through a simulation study, we demonstrate our finite-sample theoretical results for both cluster-level and individual-level treatment effects. We consider the combination of the following settings: CRTs with a large ($m = 100$) versus small ($m = 30$) number of clusters, and full-data analysis versus covariate subgroup analysis. For $i = 1, \ldots, m$, we independently generate the cluster size $N_i \sim \mathcal{U}([10, 50])$ and two cluster-level covariates $R_{i1}|N_i \sim \mathcal{N}(N_i/10, 1)$, $R_{i2}|(N_i, R_{i1}) \sim \mathcal{B}\{(1 + e^{-R_{i1}/2})^{-1}\}$, where $\mathcal{U}, \mathcal{N}, \mathcal{B}$ represent the uniform, normal, and Bernoulli distribution, respectively. For each individual $j = 1, \ldots, N_i$, we independently generate covariates $X_{ij1}|(N_i, R_{i1}, R_{i2}) \sim \mathcal{B}(0.3 + 0.4R_{i2})$ and $X_{ij2} = (2I\{R_{i1} > 0\} -$

$1)\overline{X}_{i1} + \mathcal{N}(0,1)$, and potential outcomes $Y_{ij}(a) = aN_i/50 + \sin(R_{i1})(2R_{i2} - 1) + |X_{ij1}X_{ij2}| + (1-a)\gamma_i + \varepsilon_{ij}$ for $a = 0, 1$ where $\gamma_i \sim \mathcal{N}(0, 0.5^2)$ is the random intercept and $\varepsilon_{ij} \sim N(0,1)$ is the random noise, leading to an adjusted intracluster correlation coefficient of 0.2 under $a = 0$. Then we independently generate the treatment indicator $A_i \sim \mathcal{B}(0.5)$, and obtain $Y_{ij} = A_i Y_{ij}(1) + (1 - A_i)Y_{ij}(0)$. The simulated observed data are $\{(Y_{ij}, A_i, X_{ij1}, X_{ij2}, R_{i1}, R_{i2}) : i = 1, \ldots, N_i\}_{i=1}^{m}$. To compute performance metrics, we generate 1,000 new clusters as the test data from the same data-generating distribution, and repeat the above procedure to generate 1,000 data replicates.

For each simulated data set, we first construct the conformal interval with $\alpha = 0.1$ for the cluster-average treatment effect. Given complete test data (with treatment and outcomes), we run Algorithm 1, and refer to this approach as "O". Given cluster-level covariates only, we refer to the direct approach as "B-direct" and the nested approach as "B-nested". For covariate subgroup analysis, we consider $\Omega_C = \{\overline{B}_i : R_{i1} \geq 2, R_{i2} = 1\}$, which contains 60% of all clusters. For inferring the individual-average treatment effect, we adopt the same names "O", "B-direct", "B-nested" to refer to the output of Algorithm 2, the direct approach and the nested approach, and the covariate subgroup is defined as $\Omega_I = \{B_{ij} : |X_{ij2}| < 0.5\}$,

which includes 30% of all individuals. While our theoretical results support any choice of training models, to improve predictive accuracy, we consider the training model to be an ensemble learner of linear regression and random forest implemented via the `SuperLearner` R package (van der Laan et al., 2007). We consider two metrics of performance: the probability that the conformal interval contains the true treatment effect, and the average length of the conformal interval. For the nested approach, we set $\gamma = 0.5$ to improve the informativeness of the resulting conformal interval.

## 5.2 Simulation results

Figure 1 summarizes the simulation results for the marginal and local cluster-level treatment effects given $m = 100$. In Figure 1, the upper panels show that all three methods achieve the target 90% coverage probability (reflected by the medians of all box plots sitting above 0.9), thereby confirming our theoretical results. Comparing the three methods, the "O" method has a coverage probability close to 0.9, whereas the "B-direct" method is the most conservative (coverage probability near 1). This difference can be explained by the lower panels, where the "B-direct" method yields wider conformal intervals, whose length nearly doubles the oracle length. In contrast, due to leveraging the complete information in the test data, the "O"

method achieves the near-optimal length of the conformal interval. The performance of the "B-nested" method lies between the other two since we set a loose parameter $\gamma = 0.5$. If we use $\gamma = 0.1$ instead, it will perform similarly to the "B-direct" method as demonstrated in Lei and Candès (2021) for non-clustered data. Comparing results for marginal versus local treatment effects, the coverage probability and length of conformal intervals are generally similar. In Supplementary Material D, we reproduce Figure 1 with $m = 500$, where the span of the boxplot substantially decreases and the "O" method is nearly optimal under both metrics.

Figure 2 presents the results for the marginal and local individual-level treatment effects given $m = 100$. All three methods reach the target coverage probability, with patterns similar to Figure 1. However, the conformal intervals for the individual-level treatment effect appear more stable than the cluster-level treatment effect, as reflected by shorter spans of the boxplot, but have wider lengths due to the increased variance in treatment effects. Results for $m = 500$ are presented in Supplementary Material D with similar patterns.

In Supplementary Material D, we provide additional simulations under $m = 30$ to test the performance under a smaller number of clusters. Under this scenario, we still observe valid coverage probability by the "O" and

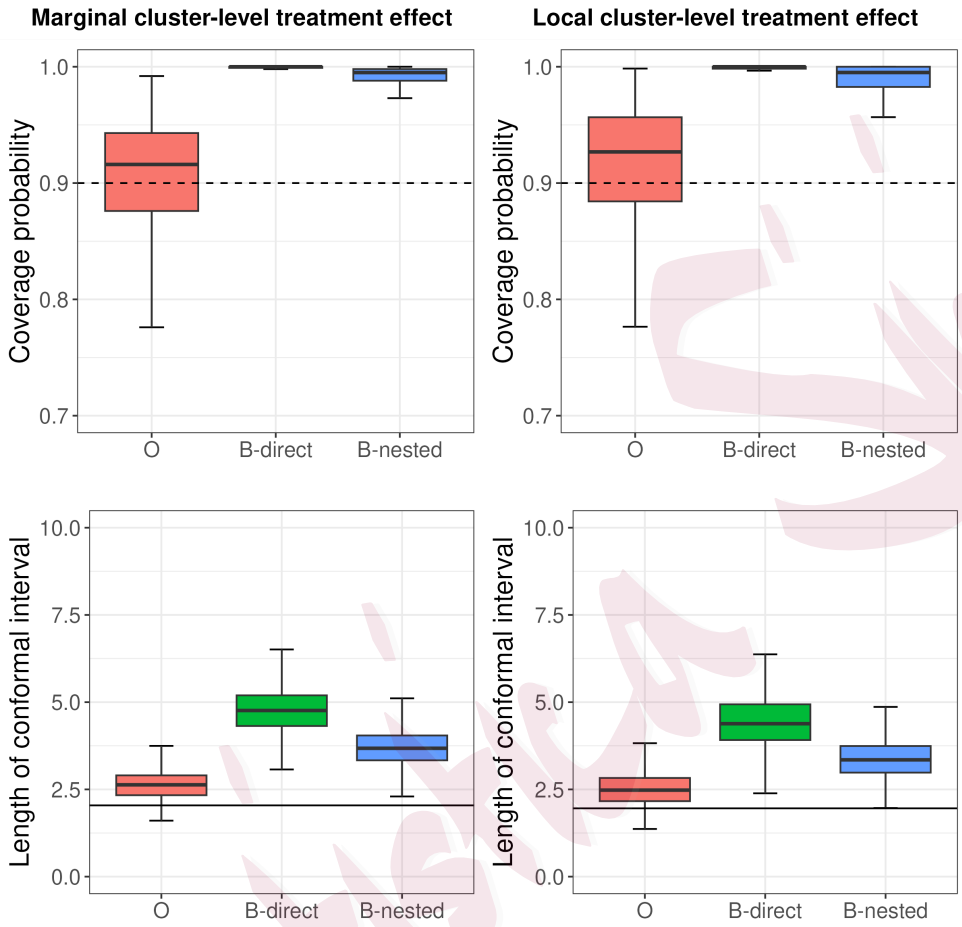


Figure 1: Simulation results (boxplot) for the marginal (left column) and local (right column, conditioning on $\{R_{i1} \geq 2, R_{i2} = 1\}$) cluster-level treatment effects with $m = 100$. In the upper panels, the dashed line is the target 90% coverage probability. In the lower panels, the solid line is the oracle length of conformal intervals, computed as the average length between the $(\alpha/2, 1-\alpha/2)$-quantiles of $\overline{Y}(1) - \overline{Y}(0)$ among test data. Each box plot is based on 1000 data points, with each representing the performance metric computed from one data replicate (instead of one test data point).

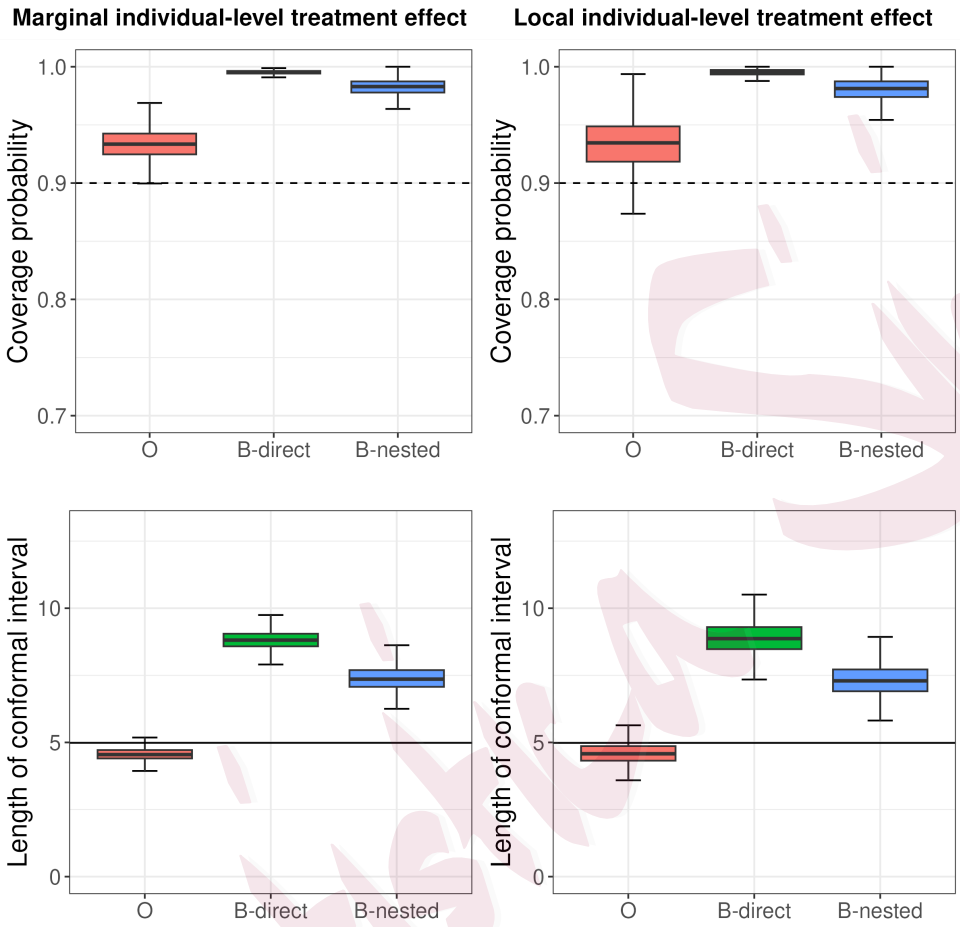


Figure 2: Simulation results (boxplot) for the marginal (left column) and local (right column, conditioning on $\{|X_{ij2}| < 0.5\}$) individual-level treatment effects with $m = 100$. In the upper panels, the dashed line is the target 90% coverage probability. In the lower panels, the solid line is the oracle length of conformal intervals, computed as the average length between the $(\alpha/2, 1-\alpha/2)$-quantiles of $Y(1) - Y(0)$ among test data.Each box plot is based on 1000 data points, with each representing the performance metric computed from one data replicate (instead of one test data point).

"B-direct" methods, but they become more conservative, as reflected by the increased length of intervals. In addition, we repeat our simulations for $m = 30$ with only linear regression as the training model (in contrast to the ensemble method). We find that, by including random forest in the ensemble method for model training, the coverage probability for the conformal intervals has negligible differences, but the length of intervals is reduced by 8.2-46.5%. This example demonstrates the improved accuracy in conformal causal inference by leveraging data-adaptive machine learners.

## 6. Data example with the PPACT cluster randomized trial

The Pain Program for Active Coping and Training study (PPACT) is a CRT evaluating the effect of a care-based cognitive behavioral therapy intervention for treating long-term opioid users with chronic pain (DeBar et al., 2022). The study equally randomized 106 primary care providers (clusters) to receive the intervention or usual care, with 1-10 participants in each cluster. We focus on the primary outcome, the PEGS (pain intensity and interference with enjoyment of life, general activity, and sleep) score at 12 months, a continuous measure of pain scale ranging from 1 to 10. For more accurate conformal causal inference, we adjust for 13 individual-level baseline variables, including the baseline PEGS score, age, gender, disability,

smoking status, body mass index, alcohol abuse, drug abuse, comorbidity, depression, number of pain types, average morphine dose, and heavy opioid usage.

In the real-world setting, we can use all 106 clusters to construct conformal interval functions $\widetilde{\mathrm{C}}_C(\overline{O})$ and $\widetilde{\mathrm{C}}_I(O)$ that are applicable to any new cluster or individuals in the new cluster by plugging in $\overline{O}_{\text{test}}$ and $O_{\text{test}}$. Here, to demonstrate our approaches, we randomly sample 20 clusters as the test data to compute performance metrics and use the rest 86 clusters to construct conformal interval functions; we repeat this process for 100 times to account for the uncertainty in the data split. We report the average and standard error for two performance metrics: length of intervals and fraction of negatives. Here, the fraction of negatives defines the proportion of conformal intervals that are subsets of $(-\infty, 0)$ among the test data. Since negative values are in the direction of treatment benefits, this metric reveals how many clusters/individuals are associated with beneficial treatment effects with probability $1-\alpha$, and bears a similar interpretation to power for a one-sided test. Because the test data have the complete information, we directly run the "O" method, i.e., Algorithm 1 and Algorithm 2, with $f_a$ set as the ensemble learner of linear regression and random forest.

Table 1 summarizes the results for the marginal treatment effects set-

ting $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$. As $\alpha$ increases, the length of intervals decreases, and the fraction of negative becomes larger. Since the treatment effect is small (relative to the variability of treatment effects, Wang et al., 2024), only a small to moderate proportion of the population has negative conformal intervals. In practice, these negative conformal intervals can be informative for new patients and clusters from the same source population generating the observed trial sample. Finally, we performed subgroup analyses on individuals with severe or moderate baseline pain, and the results are summarized in Supplementary Material D.

Table 1: Summary results of data application for marginal treatment effects. For both the length of intervals and the fraction of negatives, we present the average and standard error over 100 runs.

| Coverage probability | Marginal cluster-level treatment effect | | Marginal individual-level treatment effect | |
|---|---|---|---|---|
| | Length of intervals | Fraction of negatives | Length of intervals | Fraction of negatives |
| 90% | 4.056(0.557) | 0.089(0.069) | 6.908(0.590) | 0.055(0.026) |
| 80% | 2.874(0.412) | 0.173(0.092) | 4.800(0.345) | 0.132(0.044) |
| 70% | 2.233(0.313) | 0.238(0.104) | 3.811(0.220) | 0.199(0.052) |
| 60% | 1.799(0.255) | 0.304(0.117) | 3.108(0.189) | 0.257(0.055) |

If the inferential goal is to construct conformal intervals for a cluster or individual within the CRT sample, the target cluster can be treated as the test unit, with the remaining clusters used for training and calibration. As an illustration, Figure 3 presents 90% conformal intervals for both the cluster-level and individual-level treatment effects for a control cluster. The results indicate that, with at least 90% probability, the intervention im-

proves the mean outcome at the cluster level, as reflected by the cluster-level conformal interval (-5.01,-1.29). In contrast, none of the three individual-level conformal intervals excludes zero, suggesting insufficient evidence to support an individual-level treatment effect for this cluster.

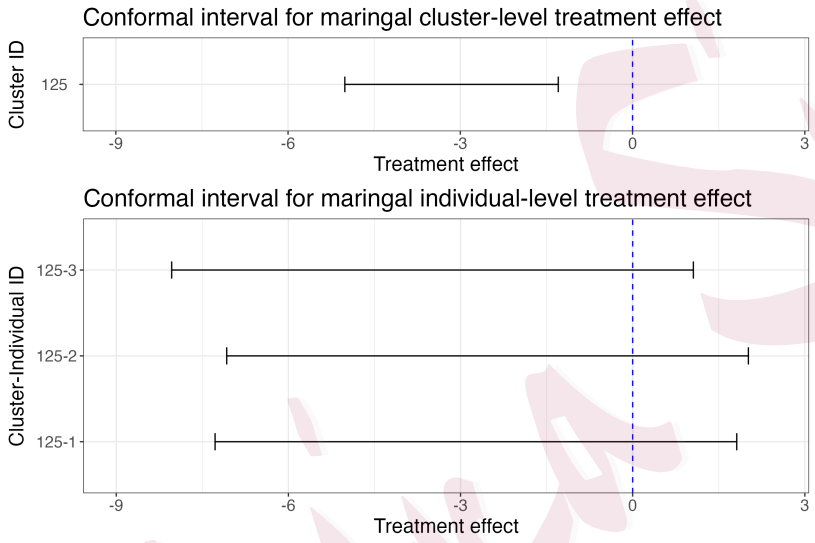


Figure 3: Cluster-level and individual-level treatment effects for one cluster with three individuals in the controlled arm of the PPACT study.

## 7. Discussion

In this article, we develop the conformal causal inference framework to study treatment effects in CRTs, and offer a complementary non-asymptotic framework to existing approaches that target the average treatment effects (e.g., Benitez et al., 2023; Su and Ding, 2021; Wang et al., 2024). Al-

though our data example in Section 6 includes a large number of clusters, our proposed methods are not confined to a large number of clusters, given appropriate adjustments of the target coverage probability. As a practical consideration, for achieving 90% coverage probability of the conformal interval for the cluster-level treatment effects, we need at least 10 clusters per arm for numerically stable calibration, and our simulation shows that the conformal interval is valid but moderately conservative given $m = 30$ and $\pi = 0.5$. Given fewer clusters, e.g., $m = 20$, targeting 90% coverage probability will only result in intervals spanning the entire real line, and more informative intervals are possible with a lower target coverage probability (say 80%). In practice, we recommend setting $\alpha = 0.05$ if $m \geq 80$, $\alpha = 0.1$ if $40 \leq m < 80$, and $\alpha = 0.2$ if $20 \leq m < 40$ given $\pi = 0.5$. If $10 < m < 20$, we can still construct the 80% conformal interval for the individual-level treatment effect, but it may be challenging to make conformal causal inference given $m \leq 10$.

The conformal framework considered here admits several reasonable variants, and our proposal reflects one specific set of design choices. First, although we establish both marginal and local coverage guarantees, the most appropriate inferential target in practice depends on the scientific question and the available data. Second, our algorithms achieve local cov-

erage through subsetting observations that satisfy the conditioning criterion. For example, one may target coverage conditional on a specific group size by restricting attention to clusters with the corresponding realized size. An alternative is weighted conformal prediction (Tibshirani et al., 2019), which uses all calibration observations by reweighting them toward the target subgroup. Compared with subsetting, this approach can improve efficiency when the relevant subset is small. However, it requires estimation of the weight function, and its theoretical guarantees typically depend on the accuracy of that estimation, with consistent estimators yielding asymptotic coverage validity. Third, although our cluster-level analysis uses the cluster-average covariate, a natural choice to predict cluster-average outcomes (Wang et al., 2026; Su and Ding, 2021), the same framework extends more generally to measurable symmetric functions $G_{N_i}(X_{i1}, \ldots, X_{iN_i})$, such as cluster totals or quantiles, for summarizing within-cluster covariates. A more systematic comparison of these design choices is an important direction for future work.

We have considered the basic CRT setting in this development, but our results can be extended in several directions. First, although we assume cluster randomization, conformal causal inference can be extended to clustered observational studies with a cluster-level treatment under strong ig-

norability, along the lines of Lei and Candès (2021) and Yang et al. (2024). With this change, one can perform conformal causal inference for treatment effects among the treated clusters and individuals, and the details are provided in Supplementary Material E. Second, we assume simple cluster randomization, whereas in other cases covariate-adaptive randomization, such as pair-matching (Balzer et al., 2016), may be used. The extension to covariate-adaptive randomization represents an important future research direction. Lastly, it would also be useful to relax the within-cluster exchangeability assumption for inferring individual-level treatment effects. Despite recent efforts (Tibshirani et al., 2019; Barber et al., 2023; Dobriban and Yu, 2025) to address this issue, non-exchangeability in clustered data still poses nontrivial challenges and warrants further investigation.

## Supplementary Materials

The Supplementary Materials include Web Appendices, Tables, and Figures, and code referenced in Sections 3-6.

## Acknowledgements

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health

under Award Number R00AI173395. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

Alaa, A. M., Z. Ahmad, and M. van der Laan (2023). Conformal meta-learners for predictive inference of individual treatment effects. *Advances in neural information processing systems 36*, 47682–47703.

Balzer, L. B., M. L. Petersen, M. J. van der Laan, and S. Collaboration (2016). Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Statistics in Medicine 35*(21), 3717–3732.

Balzer, L. B., M. L. Petersen, M. J. van der Laan, and S. Consortium (2015). Adaptive pair-matching in randomized trials with unbiased and efficient effect estimation. *Statistics in medicine 34*(6), 999–1011.

Balzer, L. B., M. van der Laan, J. Ayieko, M. Kamya, G. Chamie, J. Schwab, D. V. Havlir, and M. L. Petersen (2023). Two-stage TMLE to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics 24*(2), 502–517.

Balzer, L. B., M. J. van der Laan, M. L. Petersen, and S. Collaboration (2016). Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in medicine 35*(25), 4528–4545.

# REFERENCES

Barber, R. F., E. J. Candès, A. Ramdas, and R. J. Tibshirani (2021). Predictive inference with the jackknife+. *The Annals of Statistics 49*(1), 486 – 507.

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics 51*(2), 816–845.

Benitez, A., M. L. Petersen, M. J. van der Laan, N. Santos, E. Butrick, D. Walker, R. Ghosh, P. Otieno, P. Waiswa, and L. B. Balzer (2023). Defining and estimating effects in cluster randomized trials: A methods comparison. *Statistics in Medicine 42*(19), 3443–3466.

Breiman, L. (2001). Random forests. *Machine learning 45*, 5–32.

DeBar, L., M. Mayhew, L. Benes, A. Bonifay, R. A. Deyo, C. R. Elder, F. J. Keefe, M. C. Leo, C. McMullen, A. Owen-Smith, et al. (2022). A primary care–based cognitive behavioral therapy intervention for long-term opioid users with chronic pain: a randomized pragmatic trial. *Annals of Internal Medicine 175*(1), 46–55.

Ding, P. and L. Keele (2018). Rank tests in unmatched clustered randomized trials applied to a study of teacher training. *The Annals of Applied Statistics 12*(4), 2151–2174.

Dobriban, E. and M. Yu (2025). Symmpi: predictive inference for data with group symmetries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf022.

Dunn, R., L. Wasserman, and A. Ramdas (2023). Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association 118*(544), 2491–2502.

Hayes, R. J. and L. H. Moulton (2017). *Cluster randomised trials*. Chapman and Hall/CRC.

Hore, R. and R. F. Barber (2025). Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(2), 549–578.

Jin, Y., Z. Ren, and E. J. Candès (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences 120*(6), e2214889120.

Kahan, B. C., F. Li, A. J. Copas, and M. O. Harhay (2023). Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology 52*(1), 107–118.

Lee, Y., R. Barber, and R. Willett (2023). Distribution-free inference with hierarchical data. *ACM Journal of Data Science*.

Lei, L. and E. J. Candès (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(5), 911–938.

Li, F., J. Tong, X. Fang, C. Cheng, B. C. Kahan, and B. Wang (2025). Model-robust standardization in cluster-randomized trials. *Statistics in Medicine 44*(20-22), e70270.

Murray, D. M. et al. (1998). *Design and Analysis of Group-Randomized Trials*, Volume 29. Oxford University Press, USA.

Papadopoulos, H., K. Proedrou, V. Vovk, and A. Gammerman (2002). Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference*

*on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 345–356. Springer.

Qiu, H., E. Dobriban, and E. Tchetgen Tchetgen (2023). Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(5), 1680–1705.

Rabideau, D. J. and R. Wang (2021). Randomization-based confidence intervals for cluster randomized trials. *Biostatistics 22*(4), 913–927.

Small, D. S., T. R. Ten Have, and P. R. Rosenbaum (2008). Randomization inference in a group–randomized trial of treatments for depression: covariate adjustment, noncompliance, and quantile effects. *Journal of the American Statistical Association 103*(481), 271–279.

Su, F. and P. Ding (2021). Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society, Series B 83*(5), 994–1015.

Tibshirani, R. J., R. Foygel Barber, E. Candes, and A. Ramdas (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems 32*, 1–11.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(1), 1–21.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world*,

Volume 29. Springer.

Wang, B., M. O. Harhay, J. Tong, D. S. Small, T. P. Morris, and F. Li (2026). On the mixed-model analysis of covariance in cluster-randomized trials. *Statistical science 41*(1), 49.

Wang, B., C. Park, D. S. Small, and F. Li (2024). Model-robust and efficient covariate adjustment for cluster-randomized experiments. *Journal of the American Statistical Association 119*(548), 2959–2971.

Wang, X., K. S. Goldfeld, M. Taljaard, and F. Li (2024). Sample size requirements to test subgroup-specific treatment effects in cluster-randomized trials. *Prevention Science 25*(Suppl 3), 356–370.

Wu, J. and P. Ding (2021). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association 116*(536), 1898–1913.

Yang, Y., A. K. Kuchibhotla, and E. Tchetgen Tchetgen (2024). Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology 86*(4), 943–965.

Yin, M., C. Shi, Y. Wang, and D. M. Blei (2024). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association 119*(545), 122–135.