Supplementary Materials for "Conformal causal inference for cluster randomized trials: model-robust inference without asymptotic approximations" by Wang, Li, and Yu

In Section A, we provide additional discussion on our assumptions, including an example class of data generating distributions that imply within-cluster echangeability and a comparison with alternative assumptions in Lee et al. (2023). In Section B, we provide the proofs for our theoretical results. In Section C, we provide the nested approach for constructing conformal intervals based on CRTs. In Section D, we provide additional numerical results. In Section E, we show how the proposed approach can be extended to study the treatment effect on the treated in cluster observational studies. The R code for reproducing all simulation and data analysis is available at https://github.com/BingkaiWang/CRT-conformal.

A Additional discussion on assumptions

A.1 A class of data-generating processes that imply Assumption 3

To contextualize the within-cluster exchangeability assumption (required for inferring the individual-level treatment effect target), we discuss am example class of data-generating processes for which Assumption 3 holds. This example is meant to illustrate that, under Assumption 3, the unknown intracluster correlation structure among the outcomes can still flexibly depend on the covariates, and the model space after assuming within-cluster exchangeability remains considerably large.

We first assume that the joint distribution of $\{X_{\bullet,1}, \ldots, X_{\bullet,N}\}$ given R, N is invariant to permutation of individual component indices. A necessary condition for this process is that the marginal distribution of $X_{\bullet,j}$ is common across j. This assumption, although strong, is often invoked in simulation studies for validating and comparing methods for CRTs. Next, we consider the following restricted moment random coefficient model for the pair of individual potential outcomes:

$$\begin{pmatrix} Y_{\bullet,j}(1) \\ Y_{\bullet,j}(0) \end{pmatrix} = \begin{pmatrix} \eta_1(X_{\bullet,j}, R, N) \\ \eta_0(X_{\bullet,j}, R, N) \end{pmatrix} + \begin{pmatrix} g_1^\top(X_{\bullet,j})\gamma(1) \\ g_0^\top(X_{\bullet,j})\gamma(0) \end{pmatrix} + \begin{pmatrix} h_1(X_{\bullet,j}, R, N)\epsilon_{\bullet,j}(1) \\ h_0(X_{\bullet,j}, R, N)\epsilon_{\bullet,j}(0) \end{pmatrix}, \tag{1}$$

where η_1, η_2 are completely unspecified functions of the baseline covariates and represent fixed effects, $g_1(X_{\bullet,j})$, $g_0(X_{\bullet,j})$ are design vectors (including intercept) for random intercept and coefficients and are an unknown function of $X_{\bullet,j}$, $(\gamma^\top(1), \gamma^\top(0))^\top$ is a random effect vector that jointly follows a multivariate distribution with mean zero and variance-covariance matrix $\Lambda(R,N) = \begin{pmatrix} \Lambda_1(R,N) & \Lambda_{10}(R,N) \\ \Lambda_{10}^\top(R,N) & \Lambda_0(R,N) \end{pmatrix}$ that is allowed to depend on cluster-level covariates R, N, and independently, $(\epsilon_{\bullet,j}(1), \epsilon_{\bullet,j}(0))^\top$ are bivariate errors that follow a bivariate distribution with mean zero and variance-covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{10} \\ \sigma_{10} & \sigma_0^2 \end{pmatrix}$ and that are also independent across j. $\Sigma(X_{\bullet,j}) = \begin{pmatrix} \sigma_1^2(X_{\bullet,j}) & \sigma_{10}(X_{\bullet,j}) \\ \sigma_{10}(X_{\bullet,j}) & \sigma_0^2(X_{\bullet,j}) \end{pmatrix}$ that is allowed to depend on individual-level covariates. This data-generating process has the following features. (I) *semiparametric*: the class of data-generating processes only impose moment conditions for the cluster-level random effects (intercept and coefficients) and the individual-level random errors, but leaves the fixed-effects components and random-effects design vectors completely unspecified; (II) *heteroscedastic*: the variance-covariance matrix of the cluster-level random effects is allowed to depend on cluster-level covariates, and the variance-covariance matrix of the individual-level errors is allowed to depend on individual-level covariates, without specifying the form of dependency. When $g_1(X_{\bullet,j}) = g_0(X_{\bullet,j}) = 1$, and the distributions of

$(\gamma(1), \gamma(0))^\top$, $(\epsilon_{\bullet,j}(1), \epsilon_{\bullet,j}(0))^\top$ are both bivariate normal with constant variance-covariance matrices, (1) becomes the bivariate linear mixed model discussed, for example, in Yang et al. (2023) for CRTs with multivariate outcomes; (III) *heterogeneous intracluster correlation structure*: the class of models (1) imply the following intracluster correlation coefficients (ICCs) that can vary by pairs of individuals and by clusters, depending on the individual-level and cluster-level covariates.

1. For two treated (control) potential outcomes from different individuals in the same cluster, the *single-world between-individual ICC* is given by

$$Corr(Y_{\bullet,j}(1), Y_{\bullet,l}(1)|X_{\bullet,j}, X_{\bullet,l}, R, N) = \frac{g_1^\top(X_{\bullet,j})\Lambda_1(R,N)g_1(X_{\bullet,l})}{\sqrt{V_{\bullet,j}(1)}\sqrt{V_{\bullet,l}(1)}},$$
$$Corr(Y_{\bullet,j}(0), Y_{\bullet,l}(0)|X_{\bullet,j}, X_{\bullet,l}, R, N) = \frac{g_0^\top(X_{\bullet,j})\Lambda_0(R,N)g_0(X_{\bullet,l})}{\sqrt{V_{\bullet,j}(0)}\sqrt{V_{\bullet,l}(0)}},$$

   where $V_{\bullet,j}(a) = g_0^\top(X_{\bullet,j})\Lambda_a(R,N)g_a(X_{\bullet,j}) + h_a^2(X_{\bullet,j}, R, N)\sigma_a^2$ for $a = 0, 1$.

2. For each individual, the *cross-world within-individual ICC* for the pair of potential outcomes is

$$Corr(Y_{\bullet,j}(1), Y_{\bullet,j}(0)|X_{\bullet,j}, R, N) = \frac{g_1^\top(X_{\bullet,j})\Lambda_{10}(R,N)g_0(X_{\bullet,j}) + h_1(X_{\bullet,j}, R, N)h_0(X_{\bullet,j}, R, N)\sigma_{10}}{\sqrt{V_{\bullet,j}(1)}\sqrt{V_{\bullet,j}(0)}}.$$

3. For potential outcomes from two different individuals in the same cluster, the *cross-world between-individual ICC* is

$$Corr(Y_{\bullet,j}(1), Y_{\bullet,l}(0)|X_{\bullet,j}, X_{\bullet,l}, R, N) = \frac{g_1^\top(X_{\bullet,j})\Lambda_{10}(R,N)g_0(X_{\bullet,l})}{\sqrt{V_{\bullet,j}(1)}\sqrt{V_{\bullet,l}(0)}}.$$

Next, we show that within-cluster exchangeability holds under this class of semiparametric data-generating processes. To see this, we factorize the joint distribution of the full data

vector for a cluster by

$$f(U_{\bullet,1},\ldots,U_{\bullet,N}) = \left\{\prod_{j=1}^{N} f(Y_{\bullet,j}(1),Y_{\bullet,j}(0)|\gamma(1),\gamma(0),X_{\bullet,j},R,N)\right\} \times f(\gamma(1),\gamma(0)|R,N)f(X_{\bullet,1},\ldots,X_{\bullet,N}|R,N)f(R,N).$$

Under any random permutation of indices (mapping from $\{1,\ldots,N\}$ to $\{\sigma(1),\ldots,\sigma(N)\}$), we get

$$f(U_{\bullet,\sigma(1)},\ldots,U_{\bullet,\sigma(N)}) = \left\{\prod_{j=1}^{N} f(Y_{\bullet,\sigma(j)}(1),Y_{\bullet,\sigma(j)}(0)|\gamma(1),\gamma(0),X_{\bullet,\sigma(j)},R,N)\right\} \times f(\gamma(1),\gamma(0)|R,N)f(X_{\bullet,\sigma(1)},\ldots,X_{\bullet,\sigma(N)}|R,N)f(R,N).$$

By assumption, we have $f(X_{\bullet,\sigma(1)},\ldots,X_{\bullet,\sigma(N)}|R,N) = f(X_{\bullet,1},\ldots,X_{\bullet,N}|R,N)$, and under the bivariate random coefficient model (1), we have $\prod_{j=1}^{N} f(Y_{\bullet,j}(1),Y_{\bullet,j}(0)|\gamma(1),\gamma(0),X_{\bullet,j},R,N) = \prod_{j=1}^{N} f(Y_{\bullet,\sigma(j)}(1),Y_{\bullet,\sigma(j)}(0)|\gamma(1),\gamma(0),X_{\bullet,\sigma(j)},R,N)$; thus we have $f(U_{\bullet,\sigma(1)},\ldots,U_{\bullet,\sigma(N)}) = f(U_{\bullet,1},\ldots,U_{\bullet,N})$ and within-cluster exchangeability holds.

A.2 Comparison with alternative assumptions

In Lee et al. (2023), the hierarchical sampling assumption states that, independently for each cluster $k = 1,\ldots,K$,

$$\Pi_k \sim P_{\Pi}, \qquad N_k \mid \Pi_k \sim P_{N|\Pi}, \qquad Z_{k1},\ldots,Z_{kN_k} \mid (\Pi_k,N_k) \overset{\text{iid}}{\sim} \Pi_k,$$

where N_k is the cluster size and Z_{ij} is the individual-level data. This formulation may seem more general than our Assumption 1, since each cluster is allowed to have its own latent distribution Π_k. However, after marginalizing over the latent random measure Π_k, the induced cluster-level distribution is identical across clusters. In this sense, the two formulations are equivalent at the marginal level relevant to our assumption.

To see this, let $Z_k = (Z_{k1}, \ldots, Z_{kN_k})$. Then, for any measurable set $\mathcal{A}$,

$$\begin{aligned} P(Z_k \in \mathcal{A} \mid N_k = n) &= \int P(Z_k \in \mathcal{A} \mid \Pi_k = \pi, N_k = n)\, dP_{\Pi|N=n}(\pi) \\ &= \int \pi^{\otimes n}(\mathcal{A})\, dP_{\Pi|N=n}(\pi), \end{aligned}$$

where $\pi^{\otimes n}$ denotes the n-fold product measure of π. Therefore, conditional on $N_k = n$, the random vector Z_k has the same distribution for all clusters. This is precisely the cluster-level homogeneity imposed by our Assumption 1.

B Proofs

B.1 Proof of Theorem 1

Proof of Theorem 1. The proof of Theorem 1 follows the classical split conformal causal inference (Lei and Candès, 2021), where the key difference in our method is that we only consider the data within Ω_C to obtain the local coverage.

Since $\widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}) = (-1)^{A_{\text{test}}+1}\{\overline{Y}_{\text{test}} - \widetilde{\mathrm{C}}_{C,1-A_{\text{test}}}(\overline{B}_{\text{test}})\}$ and $\overline{Y}_{\text{test}} = A_{\text{test}}\overline{Y}_{\text{test}}(1) + (1 - A_{\text{test}})\overline{Y}_{\text{test}}(0)$, we have

$$\begin{aligned} &P\left\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}) \middle| \overline{B}_{\text{test}} \in \Omega_C\right\} \\ &= \sum_{a=0}^{1} P\left\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_C(\overline{O}_{\text{test}}), A_{\text{test}} = a \middle| \overline{B}_{\text{test}} \in \Omega_C\right\} \\ &= \sum_{a=0}^{1} P\left\{\overline{Y}_{\text{test}}(a) \notin \widetilde{\mathrm{C}}_{C,a}(\overline{B}_{\text{test}}), A_{\text{test}} = 1 - a \middle| \overline{B}_{\text{test}} \in \Omega_C\right\} \\ &\leq \sum_{a=0}^{1} P\left\{\overline{Y}_{\text{test}}(a) \notin \widetilde{\mathrm{C}}_{C,a}(\overline{B}_{\text{test}}) \middle| \overline{B}_{\text{test}} \in \Omega_C\right\}. \end{aligned}$$

Therefore, it remains to show $P\left\{\overline{Y}_{\text{test}}(a) \notin \widetilde{\mathrm{C}}_{C,a}(\overline{B}_{\text{test}}) \middle| \overline{B}_{\text{test}} \in \Omega_C\right\} \leq \alpha$ for $a = 0, 1$.

Denote $\mathcal{I} = \mathcal{I}_{ca}(a) \cup \{\text{test}\}$. By Assumptions 1-2 and the additional assumption in

Theorem 1, $\{(\overline{Y}_i(a), \overline{B}_i) : i \in \mathcal{I}\}$ are i.i.d. given $\overline{B}_i \in \Omega_C$. Of note, here we use Assumption 2 to obtain $(\overline{Y}_i(a), \overline{B}_i)|(A_i = a, \overline{B}_i \in \Omega_C)$ is identically distributed as $(\overline{Y}_i(a), \overline{B}_i)|(\overline{B}_i \in \Omega_C)$ for $i \in \mathcal{I}_{ca}(a)$. Since $\widehat{f}_a$ is a function of the training data, which are independent of the calibration data and test data, then $\{s(\overline{B}_i, \overline{Y}_i(a)), i \in \mathcal{I}\}$ are also i.i.d. conditioning on the training fold and $\overline{B}_i \in \Omega_C$. Then,

$$P\left\{ s(\overline{B}_{\text{test}}, \overline{Y}_{\text{test}}(a)) \leq \text{Quantile}_{1-\alpha}\left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \delta_{s(\overline{B}_i, \overline{Y}_i(a))} \right) \middle| \overline{B}_{\text{test}} \in \Omega_C \right\} \geq 1 - \alpha.$$

Since the quantile of an empirical distribution does not decrease if we replace one sample by $+\infty$, we also have

$$P\left\{ s(\overline{B}_{\text{test}}, \overline{Y}_{\text{test}}(a)) \leq \text{Quantile}_{1-\alpha}\left(\frac{1}{|\mathcal{I}_{ca}(a)| + 1} \sum_{i \in \mathcal{I}_{ca}(a)} \delta_{s(\overline{B}_i, \overline{Y}_i(a))} + \delta_{+\infty} \right) \middle| \overline{B}_{\text{test}} \in \Omega_C \right\} \geq 1 - \alpha.$$

Since $\overline{Y}_i(a) = \overline{Y}_i$ for $i \in \mathcal{I}_{ca}(a)$, we have

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

Therefore,

$$P\left(\overline{Y}_{\text{test}}(a) \in \widetilde{\text{C}}_{C,a}(\overline{B}_{\text{test}}) \middle| \overline{B}_{\text{test}} \in \Omega_C \right) = P\left(s(\overline{B}_{\text{test}}, \overline{Y}_{\text{test}}(a)) \leq \widehat{q}_{1-\alpha}(a) \middle| \overline{B}_{\text{test}} \in \Omega_C \right) \geq 1 - \alpha,$$

which completes the proof. □

Remark 1. *In Theorem 1, the $1 - 2\alpha$ lower bound can be tight under a worst-case treatment rule. In particular, if the target treatment distribution $\widetilde{P}_{A|W,N}$ is allowed to depend on the potential outcomes, then one can construct an adversarial assignment mechanism that selects the treatment arm whose arm-specific prediction set fails to cover. To see a concrete example, assume $A_{\text{test}} = 1$ whenever $Y_{\text{test}}(0) \notin \widetilde{\text{C}}_{C,0}(\overline{B}_{\text{test}})$ and $A_{\text{test}} = 0$ whenever*

$Y_{\text{test}}(1) \notin \widetilde{\mathrm{C}}_{C,1}(\overline{B}_{\text{test}})$. *Then we have*

$$
\begin{aligned}
&P\left\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_{C}(\overline{O}_{\text{test}})\right\} \\
&= P\left\{\overline{Y}_{\text{test}}(1) \notin \widetilde{\mathrm{C}}_{C,1}(\overline{B}_{\text{test}}), A_{\text{test}} = 0\right\} + P\left\{\overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_{C,0}(\overline{B}_{\text{test}}), A_{\text{test}} = 1\right\} \\
&= P\left\{\overline{Y}_{\text{test}}(1) \notin \widetilde{\mathrm{C}}_{C,1}(\overline{B}_{\text{test}})\right\} + P\left\{\overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_{C,0}(\overline{B}_{\text{test}})\right\}.
\end{aligned}
$$

Therefore, if each arm-specific prediction interval has miscoverage probability α, the total miscoverage can be as large as 2α, showing that the $1-2\alpha$ coverage bound is tight.

B.2 Proof of Theorem 2

Lemma 1. *Denote $M_i = \sum_{j=1}^{N_i} I\{B_{ij} \in \Omega_I\}$ and $1 \le j_1 \le \cdots \le j_{M_i} \le N_i$ be the ordered indices such that $B_{ij_k} \in \Omega_I$ for $k = 1, \ldots, M_i$. Letting e_{ij_k} be a length-N_i vector with the j_k-th entry 1 and the rest zero, we define a matrix $D_i = [e_{ij_1}, \ldots, e_{ij_{M_i}}] \in \mathbb{R}^{N_i \times M_i}$ and $\widetilde{W}_i = D_i^\top W_i$, where $W_i = (V_{i1}^\top, \ldots, V_{iN_i}^\top)^\top \in \mathbb{R}^{N_i \times (p+2)}$ for $V_{ij} = (Y_{ij}(1), Y_{ij}(0), B_{ij}) \in \mathbb{R}^{p+2}$. In other words, $\widetilde{W}_i$ is the data from cluster i after filtering out individuals with $B_{ij} \notin \Omega_I$. For notational convenience, we denote the j-the row of $\widetilde{W}_i$ as $\widetilde{V}_{ij}$. Under Assumptions 1 and 3, we have*

(a) Let $\widetilde{\mathcal{W}} = \bigsqcup_{r \ge 0} \mathbb{R}^{r \times (p+2)}$, equipped with the disjoint-union Borel σ-algebra. Then $(\widetilde{W}_1, \ldots, \widetilde{W}_m)$ are independent and identically distributed as random elements of $\widetilde{\mathcal{W}}$.

(b) Within $\widetilde{W}_i$, the distribution of $(\widetilde{V}_{i1}, \ldots, \widetilde{V}_{iM_i})$ is exchangeable conditioning on M_i.

(c) Given the above results, $\widetilde{V}_{ij}$ has the same distribution as $V_{ij}|(B_{ij} \in \Omega_I)$ for any j. It implies that $E[f(\widetilde{V}_{ij}, T)] = E[f(V_{ij}, T)|B_{ij} \in \Omega_I]$ for any random variable T that is independent of $(V_{ij}, \widetilde{V}_{ij})$ and any intergrable function f.

Proof of Lemma 1. To prove (a), we observe that the matrix $\widetilde{W}_i$ is obtained from W_i by retaining exactly those rows whose baseline covariates belong to Ω_I. We first show that $(M_i, \widetilde{W}_i)$ is a measurable function of W_i. For each fixed cluster size n and each binary vector $s = (s_1, \ldots, s_n) \in \{0,1\}^n$, define the measurable event $E_{n,s} = \Big\{N_i = n,\ I(B_{i1} \in \Omega_I) =$

$s_1, \ldots, I(B_{in} \in \Omega_I) = s_n\Big\}$. On the event $E_{n,s}$, the quantity M_i is equal to $\sum_{j=1}^{n} s_j$, and $\widetilde{W}_i$ is obtained by selecting from W_i exactly those rows j for which $s_j = 1$. Hence, on $E_{n,s}$, the map $W_i \mapsto (M_i, \widetilde{W}_i)$ reduces to a coordinate projection, which is measurable. Since for each fixed n there are only finitely many patterns $s \in \{0,1\}^n$, and the events $E_{n,s}$ are measurable, it follows that $(M_i, \widetilde{W}_i)$ is a measurable transformation of W_i. Under Assumption 1, the cluster-level variables $W_1, \ldots, W_m$ are independent and identically distributed. Because measurable transformations preserve independence and identical distribution, we conclude that $(M_1, \widetilde{W}_1), \ldots, (M_m, \widetilde{W}_m)$ are independent and identically distributed.

To prove (b), conditional on N_i, we define $Q_i = (I\{B_{i1} \in \Omega_I\}, \ldots, I\{B_{iN_i} \in \Omega_I\})$. We first prove that $(\widetilde{V}_{i1}, \ldots, \widetilde{V}_{iM_i})|Q_i = q, M_i = r$ are exchangeable.

Without loss of generality, we assume $B_{ij}, j = 1, \cdots, r \in \Omega_I$ (the first r elements). For any measureable sets $A_1, \cdots, A_r$, the conditional distribution of $(\widetilde{V}_{i1}, \ldots, \widetilde{V}_{iM_i})|Q_i = q, M_i = r$ is given by

$$(i) := P(V_{i1} \in A_1, \cdots, V_{ir} \in A_r | V_{i1} \in B, \cdots, V_{ir} \in B, V_{i(r+1)} \in C, \cdots, V_{iN_i} \in C).$$

Here we denote the domain of $(Y(1), Y(0))$ as $\mathcal{D}_Y \in \mathbb{R}^2$ and let $B = \mathcal{D}_Y \times \Omega_I$ and $C = \mathcal{D}_Y \times \Omega_I^c$. It holds that

$$(i) = \frac{P(V_{i1} \in A_1 \cap B, \cdots, V_{ir} \in A_r \cap B, V_{i(r+1)} \in C, \cdots, V_{iN_i} \in C)}{P(V_{i1} \in B, \cdots, V_{ir} \in B, V_{i(r+1)} \in C, \cdots, V_{iN_i} \in C)}$$

Since conditional on N_i, $(V_{i1}, \cdots, V_{iN_i})$ are exchangeable, for any permutation map $\sigma \in S_r$ (where S_r is the permutation group on $1, \ldots, r$), we obtain

$$\begin{aligned}
(i) &= \frac{P(V_{i\sigma(1)} \in A_1 \cap B, \cdots, V_{i\sigma(r)} \in A_r \cap B, V_{i(r+1)} \in C, \cdots, V_{iN_i} \in C)}{P(V_{i1} \in B, \cdots, V_{ir} \in B, V_{i(r+1)} \in C)} \\
&= \frac{P(V_{i\sigma(1)} \in A_1, \cdots, V_{i\sigma(r)} \in A_r, V_{i1} \in B, \cdots, V_{ir} \in B, V_{i(r+1)} \in C, \cdots, V_{iN_i} \in C)}{P(V_{i1} \in B, \cdots, V_{ir} \in B, V_{i(r+1)} \in C)} \\
&= P(V_{i\sigma(1)} \in A_1, \cdots, V_{i\sigma(r)} \in A_r | V_{i1} \in B, \cdots, V_{ir} \in B, V_{i(r+1)} \in C, \cdots, V_{iN_i} \in C).
\end{aligned}$$

Therefore, we have $[(\widetilde{V}_{i1},\ldots,\widetilde{V}_{iM_i})|Q_i=q,M_i=r] \stackrel{d}{=} [(\widetilde{V}_{i\sigma(1)},\ldots,\widetilde{V}_{i\sigma(M_i)})|Q_i=q,M_i=r]$ for any $\sigma \in S_r$.

We next prove the exchangeability by only conditioning on $M_i = r$. We know that for any $\sigma \in S_r$

$$\begin{aligned}
&P(\widetilde{V}_{i\sigma(1)},\ldots,\widetilde{V}_{i\sigma(M_i)} \in A_1 \times \cdots \times A_r | M_i = r) \\
&\quad = \sum_{Q_i=q} P(\widetilde{V}_{i\sigma(1)},\ldots,\widetilde{V}_{i\sigma(M_i)} \in A_1 \times \cdots \times A_m | Q_i = q, M_i = r) \cdot P(Q_i = q | M_i = r) \\
&\quad = \sum_{Q_i=q} P(\widetilde{V}_{i1},\ldots,\widetilde{V}_{iM_i} \in A_1 \times \cdots \times A_r | Q_i = q, M_i = r) \cdot P(Q_i = q | M_i = r) \\
&\quad = P(\widetilde{V}_{i1},\ldots,\widetilde{V}_{iM_i} \in A_1 \times \cdots \times A_r | M_i = r).
\end{aligned}$$

We then conclude our proof of (b).

Finally, we prove (c). By definition, $\widetilde{V}_{ij}$ is equal to $V_{ij'}$ for some j' such that $B_{ij'} \in \Omega_I$. Therefore, $\widetilde{V}_{ij} \stackrel{d}{=} V_{ij'}|(B_{ij'} \in \Omega_I)$. By the exchangeability of W_i, we have $\widetilde{V}_{ij} \stackrel{d}{=} V_{ij}|(B_{ij} \in \Omega_I)$. In addition,

$$\begin{aligned}
E[f(V_{ij},T)|B_{ij} \in \Omega_I] &= \int f(V_{ij},T) dP(V_{ij},T|B_{ij} \in \Omega_I) \\
&= \int f(V_{ij},T) d\{P(V_{ij}|B_{ij} \in \Omega_I)P(T)\} \\
&= \int f(\widetilde{V}_{ij},T) d\{P(\widetilde{V}_{ij})P(T)\} \\
&= \int f(\widetilde{V}_{ij},T) dP(\widetilde{V}_{ij},T) \\
&= E[f(\widetilde{V}_{ij},T)],
\end{aligned}$$

where the second and fourth equation uses the independence between T and $(V_{ij}, \widetilde{V}_{ij})$, and the thrid equation results from $\widetilde{V}_{ij} \stackrel{d}{=} V_{ij}|(B_{ij} \in \Omega_I)$. □

Proof of Theorem 2. Since $\widetilde{\mathrm{C}}_I(O_{\text{test}}) = (-1)^{A_{\text{test}}+1}\{Y_{\text{test}} - \widetilde{\mathrm{C}}_{I,1-A_{\text{test}}}(B_{\text{test}})\}$ and $Y_{\text{test}} = A_{\text{test}}Y_{\text{test}}(1) +$

$(1 - A_{\text{test}})Y_{\text{test}}(0)$, we have

$$
\begin{aligned}
&P\left\{ Y_{\text{test}}(1) - Y_{\text{test}}(0) \notin \widetilde{\text{C}}_I(O_{\text{test}}) \middle| B_{\text{test}} \in \Omega_C \right\} \\
&= \sum_{a=0}^{1} P\left\{ Y_{\text{test}}(1) - Y_{\text{test}}(0) \notin \widetilde{\text{C}}_I(O_{\text{test}}), A_{\text{test}} = a \middle| B_{\text{test}} \in \Omega_C \right\} \\
&= \sum_{a=0}^{1} P\left\{ Y_{\text{test}}(a) \notin \widetilde{\text{C}}_{I,a}(B_{\text{test}}), A_{\text{test}} = 1 - a \middle| B_{\text{test}} \in \Omega_C \right\} \\
&\leq \sum_{a=0}^{1} P\left\{ Y_{\text{test}}(a) \notin \widetilde{\text{C}}_{I,a}(B_{\text{test}}) \middle| B_{\text{test}} \in \Omega_C \right\}.
\end{aligned}
$$

Therefore, it remains to show $P\left\{ Y_{\text{test}}(a) \notin \widetilde{\text{C}}_{I,a}(B_{\text{test}}) \middle| B_{\text{test}} \in \Omega_C \right\} \leq \alpha$ for $a = 0, 1$.

To this end, we denote W_{m+1} as a new cluster independently sampled from $\mathcal{P}^W$. By construction, $(Y_{\text{test}}(1), Y_{\text{test}}(0), B_{\text{test}})$ comes from an arbitrary individual in this cluster. By Assumption 3, $(Y_{\text{test}}(1), Y_{\text{test}}(0), B_{\text{test}})$ is identically distributed as $V_{m+1,1} = (Y_{m+1,1}(1), Y_{m+1,1}(0), B_{m+1,1})$, where $V_{m+1,j}$ is the j-th row of W_{m+1}. Therefore, our goal is to show $P\left\{ Y_{m+1,1}(a) \notin \widetilde{\text{C}}_{I,a}(B_{m+1,1}) \middle| B_{m+1,1} \in \Omega_C \right\} \leq \alpha$ for $a = 0, 1$. To further simplify the goal, we denote $\widetilde{W}_{m+1}$ as the new cluster after we filter out individuals with $B_{m+1,j} \notin \Omega_I$ and let $\widetilde{V}_{m+1,j} = (\widetilde{Y}_{m+1,j}(1), \widetilde{Y}_{m+1,j}(0), \widetilde{B}_{m+1,j})$ denote the j-th row of $\widetilde{W}_{m+1}$. (See Lemma 1 for the detailed construction for $\widetilde{W}_{m+1}$). By Lemma 1 (c), $\widetilde{V}_{m+1,1}$ is identically distributed as $V_{m+1,1} | (B_{m+1,1} \in \Omega_I)$. Since $\widetilde{\text{C}}_{I,a}$ is a function of $\{W_i : i = 1, \ldots, m, A_i = a\}$, which are independent of W_{m+1}, Lemma 1(c) further implies

$$
P\left\{ Y_{m+1,1}(a) \notin \widetilde{\text{C}}_{I,a}(B_{m+1,1}) \middle| B_{m+1,1} \in \Omega_C \right\} = P\left\{ \widetilde{Y}_{m+1,1}(a) \notin \widetilde{\text{C}}_{I,a}(\widetilde{B}_{m+1,1}) \right\}.
$$

This result yields our final goal to show $P\left\{ \widetilde{Y}_{m+1,1}(a) \notin \widetilde{\text{C}}_{I,a}(\widetilde{B}_{m+1,1}) \right\} \leq \alpha$.

Next, we denote the calibration data $\mathcal{O}_{ca}(a) = \{\widetilde{W}_i : i \in \mathcal{I}_{ca}(a)\}$. Like $\widetilde{W}_{m+1}$, $\widetilde{W}_i$ contains those individuals in W_i with $B_{ij} \in \Omega_I$. By Assumptions 1-2 and Lemma 1(a), $\{\widetilde{W}_i :\in \mathcal{I}\}$ are i.i.d., where $\mathcal{I} = \mathcal{I}_{ca}(a) \cup \{m+1\}$. Within each $\widetilde{W}_i$, Lemma 1 (b) implies that $(\widetilde{V}_{i1}, \ldots, \widetilde{V}_{iM_i})$

are exchangeable conditioning on M_i within each cluster.

We define a function

$$q_{1-\alpha}(\{\widetilde{W}_i, i \in \mathcal{I}\}) = \text{Quantile}_{1-\alpha}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{M_i}\sum_{j=1}^{M_i}\delta_{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a))}\right),$$

which is the $(1-\alpha)$-quantile of the weighted empirical distribution for the calibration data and new cluster. By definition of the quantile function, we have, for any $i \in \mathcal{I}$ and $j \in \{1,\ldots,M_i\}$,

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{M_i}\sum_{j=1}^{M_i} I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, i \in \mathcal{I}\})\right\} \ge 1-\alpha.$$

Consider any permutation σ_i on $(1,\ldots,M_i)$. Since we have shown that each $\widetilde{W}_i$ is exchangeable given M_i, we have, for each $i \in \mathcal{I}$,

$$I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, i \in \mathcal{I}\})\right\} \stackrel{d}{=} I\left\{s(\widetilde{B}_{i\sigma(j)},\widetilde{Y}_{i\sigma(j)}(a)) \le q_{1-\alpha}(\{\widetilde{W}_{i'}, i' \in \mathcal{I}, i' \ne i\} \cup \{\sigma(\widetilde{W}_i)\})\right\},$$

conditioning on the training fold and M_i. Since the weighted empirical distribution is invariant to within-cluster permutations, we have $q_{1-\alpha}(\{\widetilde{W}_{i'}, i' \in \mathcal{I}, i' \ne i\} \cup \{\sigma(\widetilde{W}_i)\}) = q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})$. Combined with the fact that the training fold is independent of the calibration data and the test data, we have

$$E\left[I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\Big|M_i\right] = E\left[I\left\{s(\widetilde{B}_{i1},\widetilde{Y}_{i1}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\Big|M_i\right]$$

by choosing permutations σ with $\sigma(j) = 1$. After averaging over i and marginalizing over M_i, we have

$$E\left[I\left\{s(\widetilde{B}_{i1},\widetilde{Y}_{i1}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\right] = E\left[\frac{1}{M_i}\sum_{j=1}^{M_i} I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\right].$$

Since we have shown that $\widetilde{W}_i$ are i.i.d., we also have

$$
\begin{aligned}
&E\left[\frac{1}{M_i}\sum_{j=1}^{M_i} I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\right] \\
&= E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{M_i}\sum_{j=1}^{M_i} I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\right].
\end{aligned}
$$

Taken together, we get

$$
\begin{aligned}
&E\left[I\left\{s(\widetilde{B}_{m+1,1},\widetilde{Y}_{m+1,1}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\right] \\
&= E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{M_i}\sum_{j=1}^{M_i} I\left\{s(\widetilde{B}_{ij},\widetilde{Y}_{ij}(a)) \le q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})\right\}\right] \\
&\ge 1-\alpha.
\end{aligned}
$$

Finally, we need to connect $q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\})$ to $\widehat{q}_{1-\alpha}$ defined in Algorithm 3. Since replacing some point masses in the weighted empirical distribution by $\delta_{+\infty}$ does not decrease the $(1-\alpha)$-quantile, we have $q_{1-\alpha}(\{\widetilde{W}_i, \in \mathcal{I}\}) \le \widehat{q}_{1-\alpha}$, which implies $P(s(\widetilde{B}_{m+1,1},\widetilde{Y}_{m+1,1}(a)) \le \widehat{q}_{1-\alpha}) \ge 1-\alpha$. Since we defined $\widetilde{\mathrm{C}}_{I,a}(B) = \{y \in \mathbb{R} : |y - \widehat{f}_a(B)| \le \widehat{q}_{1-\alpha}(a)\}$, then the event $Y_{m+1,1} \in \widetilde{\mathrm{C}}_{I,a}(B_{m+1,1})$ is equal to $s(\widetilde{B}_{m+1,1},\widetilde{Y}_{m+1,1}(a)) \le \widehat{q}_{1-\alpha}$, which implies $P\left\{\widetilde{Y}_{m+1,1}(a) \notin \widetilde{\mathrm{C}}_{I,a}(\widetilde{B}_{m+1,1})\right\} \le \alpha$. This completes the proof. □

C Nested approaches to construct conformal intervals

C.1 Cluster-level treatment effect

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

Input: Cluster-level data $\{(\overline{Y}_i, A_i, \overline{B}_i) : i = 1, \ldots, m\}$, a test point $\overline{B}_{\text{test}}$, a prediction model $f_a(\overline{B})$ for $\overline{Y}(a)$, $a \in \{0, 1\}$, prediction models $\{m^L(\overline{B}), m^R(\overline{B})\}$ for the $(\alpha/2, 1-\alpha/2)$-quantiles of $\overline{Y}(1) - \overline{Y}(0)$, a covariate-subgroup of interest Ω_C, and levels (α, γ).

Procedure:

1. For $a = 0, 1$, 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. Use the training fold to run Step 1 of Algorithm 1 of the main paper and obtain the $(1-\alpha)$ conformal interval $\widetilde{\mathrm{C}}_{C,a}(\overline{B})$ for $\overline{Y}(a), a = 0, 1$.
3. For each $i = 1, \ldots, m$, define $\overline{C}_i = A_i\{\overline{Y}_i - \widetilde{\mathrm{C}}_{C,0}(\overline{B}_i)\} + (1 - A_i)\{\widetilde{\mathrm{C}}_{C,1}(\overline{B}_i) - \overline{Y}_i\}$ and denote $\overline{C}_i = [\overline{C}_i^L, \overline{C}_i^R]$.
4. Train prediction models $m^L(\overline{B})$ for $\overline{C}^L$ and $m^R(\overline{B})$ for $\overline{C}^R$ using all data in the training fold $\{\mathcal{O}_{tr}(a) : a = 0, 1\}$, and obtain the estimated models $\widehat{m}^L(\overline{B})$ and $\widehat{m}^R(\overline{B})$.
5. For each $i \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)$, compute the non-conformity score

$$s^*(\overline{B}_i, \overline{C}_i) = \max\{\widehat{m}^L(\overline{B}_i) - \overline{C}_i^L, \overline{C}_i^R - \widehat{m}^R(\overline{B}_i)\}.$$

6. Compute the $1-\gamma$ quantile $\widehat{q}^*_{1-\gamma}$ of the distribution

$$\widehat{F}^* = \frac{1}{|\mathcal{I}_{ca}(1)| + |\mathcal{I}_{ca}(0)| + 1} \left\{ \sum_{i \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)} \delta_{s^*(\overline{B}_i, \overline{C}_i)} + \delta_{+\infty} \right\}.$$

Output: $\widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}) = [\widehat{m}^L(\overline{B}_{\text{test}}) - \widehat{q}^*_{1-\gamma}, \widehat{m}^R(\overline{B}_{\text{test}}) + \widehat{q}^*_{1-\gamma}]$.

Theorem C.1. *Assume Assumptions 1-2 and that* $(\overline{Y}_{\text{test}}(1), \overline{Y}_{\text{test}}(0), \overline{B}_{\text{test}})$ *is an independent from the distribution induced by* $\mathcal{P}^W$. *Then, the* $\widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}})$ *output by Algorithm C.1 satisfies*

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

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

Proof of Theorem C.1. Following the proof of Theorem 1, we have $P\left\{\overline{Y}(a) \notin \widetilde{\mathrm{C}}_{C,a}(\overline{B}) \middle| \overline{B} \in \Omega_C\right\} \leq \alpha$ for $a = 0, 1$. Since Assumption 2 implies that A is independent of $(\overline{Y}(a), \overline{B})$, then

$$
\begin{aligned}
& P\left\{\overline{Y}(1) - \overline{Y}(0) \in A\{\overline{Y} - \widetilde{\mathrm{C}}_{C,0}(\overline{B})\} + (1-A)\{\widetilde{\mathrm{C}}_{C,1}(\overline{B}) - \overline{Y}\} \middle| \overline{B} \in \Omega_C\right\} \\
&= \sum_{a=0}^{1} P\left\{\overline{Y}(a) \in \widetilde{\mathrm{C}}_{C,a}(\overline{B}), A = a \middle| \overline{B} \in \Omega_C\right\} \\
&= \sum_{a=0}^{1} P\left\{\overline{Y}(a) \in \widetilde{\mathrm{C}}_{C,a}(\overline{B}) \middle| \overline{B} \in \Omega_C\right\} P(A = a) \\
&\geq 1 - \alpha.
\end{aligned}
$$

Therefore, by defining $\overline{C}_i = A_i\{\overline{Y}_i - \widetilde{\mathrm{C}}_{C,0}(\overline{B}_i)\} + (1 - A_i)\{\widetilde{\mathrm{C}}_{C,1}(\overline{B}_i) - \overline{Y}_i\}$, we have $\overline{C}_i = [\overline{C}_i^L, \overline{C}_i^R]$ representing the $(\alpha/2, 1-\alpha/2)$-quantile of $\overline{Y}_i(1) - \overline{Y}_i(0)$.

Next, we independently generate A_{test} from $\mathcal{P}^A$ and compute $\overline{C}_{\text{test}} = A_{\text{test}}\{\overline{Y}_{\text{test}} - \widetilde{\mathrm{C}}_{C,0}(\overline{B}_{\text{test}})\} + (1 - A_{\text{test}})\{\widetilde{\mathrm{C}}_{C,1}(\overline{B}_{\text{test}}) - \overline{Y}_{\text{test}}\}$. As a result, $(\overline{B}_{\text{test}}, \overline{C}_{\text{test}})$ is independent and identically distributed as $(\overline{B}_i, \overline{C}_i)$. We observe that $\widetilde{\mathrm{C}}_{C,a}, \widehat{m}^L, \widehat{m}^R$ are all functions of the training folds $\{\mathcal{O}_{tr}(a) : a = 0, 1\}$, which are independent of the calibration data and test data. Denoting $\mathcal{I} = \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0) \cup \{\text{test}\}$, by Assumptions 1-2, $\{s^*(\overline{B}_i, \overline{C}_i), i \in \mathcal{I}\}$ are i.i.d. conditioning on the training folds and $\overline{B}_i \in \Omega_C$. Then, we have

$$
P\left\{s^*(\overline{B}_{\text{test}}, \overline{C}_{\text{test}}) \leq \text{Quantile}_{1-\alpha}\left(\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)} \delta_{s^*(\overline{B}_i, \overline{C}_i)} + \delta_{+\infty}\right) \middle| \overline{B} \in \Omega_C\right\} \geq 1 - \gamma.
$$

Since we defined

$$
\text{Quantile}_{1-\gamma}\left(\frac{1}{|\mathcal{I}|}\sum_{i \in \mathcal{I}_{ca}(a)} \delta_{s^*(\overline{B}_i, \overline{C}_i)} + \delta_{+\infty}\right) = \text{Quantile}_{1-\gamma}(\widehat{F}^*) = \widehat{q}^*_{1-\gamma},
$$

we obtain

$$
\begin{aligned}
P\{\overline{C}_{\text{test}} \subset \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}) | \overline{B} \in \Omega_C\} &= P\left\{ \widehat{m}^L(\overline{B}_{\text{test}}) - \widehat{q}^*_{1-\gamma} \leq C^L_{\text{test}}, C^R_{\text{test}} \leq \widehat{m}^R(\overline{B}_{\text{test}}) + \widehat{q}^*_{1-\gamma} \middle| \overline{B} \in \Omega_C \right\} \\
&= P\{\max\{\widehat{m}^L(\overline{B}_{\text{test}}) - C^L_{\text{test}}, C^R_{\text{test}} - \widehat{m}^R(\overline{B}_{\text{test}})\} \leq \widehat{q}^*_{1-\gamma} | \overline{B} \in \Omega_C\} \\
&= P(s^*(\overline{B}_{\text{test}}, \overline{C}_{\text{test}}) \leq \widehat{q}^*_{1-\gamma} | \overline{B} \in \Omega_C) \\
&\geq 1 - \gamma.
\end{aligned}
$$

Combined with the fact that $P\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \in C_{\text{test}} | \overline{B} \in \Omega_C\} \geq 1 - \alpha$, we have

$$
\begin{aligned}
&P\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}) | \overline{B} \in \Omega_C\} \\
&= P\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}), \overline{C}_{\text{test}} \subset \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}) | \overline{B} \in \Omega_C\} \\
&\quad + P\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}), \overline{C}_{\text{test}} \not\subset \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}) | \overline{B} \in \Omega_C\} \\
&\leq P\{\overline{Y}_{\text{test}}(1) - \overline{Y}_{\text{test}}(0) \notin \overline{C}_{\text{test}} | \overline{B} \in \Omega_C\} + P\{\overline{C}_{\text{test}} \not\subset \widetilde{\mathrm{C}}_C(\overline{B}_{\text{test}}) | \overline{B} \in \Omega_C\} \\
&\leq \alpha + \gamma,
\end{aligned}
$$

which completes the proof.

□

C.2 Individual-level treatment effect

Theorem C.2. *Assume Assumptions 1-3 and that* $(Y_{\text{test}}(1), Y_{\text{test}}(0), B_{\text{test}})$ *is an arbitrary individual from a new cluster independently sampled from* $\mathcal{P}^W$*. Then, the* $\widetilde{\mathrm{C}}_I(B_{\text{test}})$ *output by Algorithm C.2 satisfies*

$$
P\left\{ Y_{\text{test}}(1) - Y_{\text{test}}(0) \in \widetilde{\mathrm{C}}_I(B_{\text{test}}) \middle| B_{\text{test}} \in \Omega_I \right\} \geq 1 - \alpha - \gamma \tag{3}
$$

for any set Ω_I *in the support of* B_{test} *with a positive measure.*

Proof of Theorem C.2. Following the proof of Theorem 2, we have $P(\widetilde{Y}(a) \in \widetilde{\mathrm{C}}_{I,a}(\widetilde{B})) \geq$

Algorithm C.2 Computing the conformal interval $\widetilde{\mathrm{C}}_I(B)$ for indivdiual-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 B_{test}, a prediction model $f_a(B)$ for $Y(a)$, $a \in \{0, 1\}$, prediction models $\{m^L(B), m^R(B)\}$ for the $(\alpha/2, 1-\alpha/2)$-quantiles of $Y(1)-Y(0)$, a covariate-subgroup of interest Ω_I, and levels (α, γ).

Procedure:

1. For $a = 0, 1$, randomly partition 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 $\mathcal{I}_{ca}(a)$.
2. Use the training fold to run Step 1 of Algorithm 2 of the main paper and obtain the $(1-\alpha)$ conformal interval $\widetilde{\mathrm{C}}_{I,a}(B)$ for $Y(a), a = 0, 1$.
3. For each $i = 1, \ldots, m$, define $C_{ij} = A_i\{Y_{ij} - \widetilde{\mathrm{C}}_{I,0}(B_{ij})\} + (1 - A_i)\{\widetilde{\mathrm{C}}_{I,1}(B_{ij}) - Y_{ij}\}$ and denote $C_{ij} = [C_{ij}^L, C_{ij}^R]$.
4. Train prediction models $m^L(B)$ for C^L and $m^R(B)$ for C^R using all data in the training fold $\{\mathcal{O}_{tr}(a) : a = 0, 1\}$, and obtain the estimated models $\widehat{m}^L(B)$ and $\widehat{m}^R(B)$.
5. For each $(i, j) \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)$, compute the non-conformity score

$$s^*(B_{ij}, C_{ij}) = \max\{\widehat{m}^L(B_{ij}) - C_{ij}^L, C_{ij}^R - \widehat{m}^R(B_{ij})\}.$$

6. Compute the $1-\gamma$ quantile $\widehat{q}^*_{1-\gamma}$ of the distribution

$$\widehat{F}^* = \frac{1}{|\mathcal{I}_{ca}(1)| + |\mathcal{I}_{ca}(0)| + 1} \left\{ \sum_{i \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)} \frac{1}{\sum_{j=1}^{N_i} I\{B_{ij} \in \Omega_I\}} \sum_{j=1}^{N_i} I\{B_{ij} \in \Omega_I\} \delta_{s^*(B_{ij}, C_{ij})} + \delta_{+\infty} \right\}.$$

Output: $\widetilde{\mathrm{C}}_C(B_{\text{test}}) = [\widehat{m}^L(B_{\text{test}}) - \widehat{q}^*_{1-\gamma}, \widehat{m}^R(B_{\text{test}}) + \widehat{q}^*_{1-\gamma}]$.

$1-\alpha$, where $\widetilde{Y}(a)$ and $\widetilde{B}$ represent the potential outcome $Y(a)$ and covariates B given $B \in \Omega_I$. Since Assumption 2 implies that A is independent of $(\widetilde{Y}(a), \widetilde{B}))$, then

$$\begin{aligned}
&P\left\{\widetilde{Y}(1) - \widetilde{Y}(0) \in A\{Y - \widetilde{\mathrm{C}}_{C,0}(\widetilde{B})\} + (1-A)\{\widetilde{\mathrm{C}}_{C,1}(\widetilde{B}) - \widetilde{Y}\}\right\} \\
&= \sum_{a=0}^{1} P\left\{\widetilde{Y}(a) \in \widetilde{\mathrm{C}}_{C,a}(\widetilde{B}), A = a\right\} \\
&= \sum_{a=0}^{1} P\left\{\widetilde{Y}(a) \in \widetilde{\mathrm{C}}_{C,a}(\widetilde{B})\right\} P(A = a) \\
&\geq 1 - \alpha.
\end{aligned}$$

Therefore, by defining $\widetilde{C}_{ij} = A_i\{\widetilde{Y}_{ij} - \widetilde{\mathrm{C}}_{C,0}(\widetilde{B}_{ij})\} + (1-A_i)\{\widetilde{\mathrm{C}}_{C,1}(\widetilde{B}_{ij}) - \widetilde{Y}_{ij}\}$, we have $\widetilde{C}_{ij} = [\widetilde{C}^L_{ij}, \widetilde{C}^R_{ij}]$ representing the $(\alpha/2, 1-\alpha/2)$-quantile of $\widetilde{Y}_{ij}(1) - \widetilde{Y}_{ij}(0)$.

For the new cluster, in addition to sampling potential outcomes and covariates from $\mathcal{P}^W$, we independently generate A_{m+1} from $\mathcal{P}^A$, and compute $\widetilde{C}_{m+1,j} = A_{m+1}\{Y_{m+1,j} - \widetilde{\mathrm{C}}_{C,0}(B_{m+1,j})\} + (1-A_{m+1})\{\widetilde{\mathrm{C}}_{C,1}(\widetilde{B}_{m+1,j}) - \widetilde{Y}_{m+1,j}\}$. By construction, $(\widetilde{B}_{m+1,j}, \widetilde{C}_{m+1,j})$ is independently and identically distributed as $(\widetilde{B}_{ij}, \widetilde{C}_{ij})$. We observe that $\widetilde{\mathrm{C}}_{C,a}, \widehat{m}^L, \widehat{m}^R$ are all functions of the training folds $\{\mathcal{O}_{tr}(a) : a = 0, 1\}$, which are independent of the calibration data and test data. Denoting $\mathcal{I} = \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0) \cup \{\text{test}\}$, by Assumptions 1-2, $\{s^*(\widetilde{B}_i, \widetilde{C}_i), i \in \mathcal{I}\}$ are i.i.d. conditioning on the training folds and $\widetilde{B}_i \in \Omega_C$. Then, we have

$$P\left\{ s^*(\widetilde{B}_{m+1,1}, \widetilde{C}_{m+1,1}) \le \text{Quantile}_{1-\gamma}\left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)} \frac{1}{M_i} \sum_{j=1}^{M_i} \delta_{s^*(\widetilde{B}_{ij}, \widetilde{C}_{ij})} + \delta_{+\infty} \right) \right\} \ge 1 - \gamma.$$

By the definition of $\widehat{F}^*$, we have

$$\text{Quantile}_{1-\gamma}\left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}_{ca}(1) \cup \mathcal{I}_{ca}(0)} \frac{1}{M_i} \sum_{j=1}^{M_i} \delta_{s^*(\widetilde{B}_{ij}, \widetilde{C}_{ij})} + \delta_{+\infty} \right) = \text{Quantile}_{1-\gamma}(\widehat{F}^*) = \widehat{q}^*_{1-\gamma},$$

we obtain

$$\begin{aligned}
P\{\widetilde{C}_{m+1,1} \subset \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1})\} &= P\left\{ \widehat{m}^L(\widetilde{B}_{m+1,1}) - \widehat{q}^*_{1-\gamma} \le C^L_{m+1,1}, C^R_{m+1,1} \le \widehat{m}^R(\widetilde{B}_{m+1,1}) + \widehat{q}^*_{1-\gamma} \right\} \\
&= P\{\max\{\widehat{m}^L(\widetilde{B}_{m+1,1}) - C^L_{m+1,1}, C^R_{m+1,1} - \widehat{m}^R(\widetilde{B}_{m+1,1})\} \le \widehat{q}^*_{1-\gamma}\} \\
&= P(s^*(\widetilde{B}_{m+1,1}, \widetilde{C}_{m+1,1}) \le \widehat{q}^*_{1-\gamma}) \\
&\ge 1 - \gamma.
\end{aligned}$$

Combined with the fact that $P\{\widetilde{Y}_{m+1,1}(1) - \widetilde{Y}_{m+1,1}(0) \in C_{m+1,1}\} \geq 1-\gamma$, we have

$$
\begin{aligned}
&P\{\widetilde{Y}_{m+1,1}(1) - \widetilde{Y}_{m+1,1}(0) \notin \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1})\} \\
&= P\{\widetilde{Y}_{m+1,1}(1) - \widetilde{Y}_{m+1,1}(0) \notin \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1}), \widetilde{C}_{m+1,1} \subset \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1})\} \\
&\quad + P\{\widetilde{Y}_{m+1,1}(1) - \widetilde{Y}_{m+1,1}(0) \notin \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1}), \widetilde{C}_{m+1,1} \not\subset \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1})\} \\
&\leq P\{\widetilde{Y}_{m+1,1}(1) - \widetilde{Y}_{m+1,1}(0) \notin \widetilde{C}_{m+1,1}\} + P\{\widetilde{C}_{m+1,1} \not\subset \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1})\} \\
&\leq \alpha + \gamma.
\end{aligned}
$$

Finally, by Lemma 1 (c), $P\{\widetilde{Y}_{m+1,1}(1) - \widetilde{Y}_{m+1,1}(0) \notin \widetilde{\mathrm{C}}_C(\widetilde{B}_{m+1,1})\} = P\{Y_{\text{test}}(1) - Y_{\text{test}}(0) \notin \widetilde{\mathrm{C}}_C(B_{\text{test}}) | B_{\text{test}} \in \Omega_I\}$, which completes the proof. □

D Additional numerical results

D.1 Simulations with 30 clusters

When $m = 30$, the simulation setting is the same as Section 5.1 of the main paper. We provide the simulation results in Table 1. Due to limited clusters, we do not implement the "B-nested" method (which requires two sample splittings) or perform the covariate-subgroup analysis for the cluster-level treatment effect (which drops 40% of all clusters). For the remaining scenarios, the "O" and "B-direct" methods can still achieve the target coverage probability, confirming their finite-sample validity. However, they become more conservative compared to $m = 100$, as reflected by the increased length of intervals. This is expected given limited clusters used in model training and quantile estimation. An ad-hoc solution is to consider an 80% coverage probability, in which case the resulting conformal intervals have a reduced length and are likely more informative. Finally, it is worth highlighting that, even with few clusters, we are able to perform subgroup analysis for the individual-level treatment effect, and its empirical performance is similar to the marginal analysis that uses all data. This observation confirms the statistical benefit of conformal causal inference for individual-level treatment effects in small CRTs.

Table 1: Summary of simulation results for 80% and 90% conformal intervals given $m = 30$ clusters. For both coverage probability and length of intervals, we report their averages and standard errors. The local treatment effect is conditioned on $|X_{ij1}| < 0.5$.

Treatment effects	Methods	$\alpha = 0.2$		$\alpha = 0.1$	
		Coverage probability	Length of intervals	Coverage probability	Length of intervals
Marginal cluster-level	O	0.816(0.094)	3.119(1.465)	0.912(0.072)	4.211(2.230)
	B-direct	0.998(0.008)	5.891(2.833)	1.000(0.002)	7.970(4.447)
Marginal individual-level	O	0.868(0.035)	4.311(0.489)	0.977(0.014)	6.637(0.794)
	B-direct	0.989(0.007)	8.345(0.768)	1.000(0.000)	12.905(1.232)
Local individual-level	O	0.870(0.048)	4.337(0.675)	0.977(0.021)	6.933(1.260)
	B-direct	0.988(0.009)	8.398(1.138)	1.000(0.000)	13.542(2.114)

D.2 Simulations with 500 clusters

The simulations in Section of the main paper are summarized in Figures 1 and 2 below.

D.3 Simulations with linear regression as the prediction model

The simulation results with linear regression as the prediction model is given in Table 2. We compare it with Table 1, which uses linear regression and random forest as the prediction model. We observe a short length of conformal intervals when using random forest.

Table 2: Summary of simulation results for 80% and 90% conformal intervals given $m = 30$ clusters with linear regression as the prediction model. For both coverage probability and length of intervals, we report their averages and standard errors. The local treatment effect is conditioned on $|X_{ij1}| < 0.5$.

Treatment effects	Methods	$\alpha = 0.2$		$\alpha = 0.1$	
		Coverage probability	Length of intervals	Coverage probability	Length of intervals
Marginal cluster-level	O	0.821(0.095)	5.149(4.266)	0.915(0.071)	7.457(8.898)
	B-direct	0.999(0.007)	10.310(8.526)	1.000(0.001)	14.910(8.017)
Marginal individual-level	O	0.865(0.044)	4.966(1.114)	0.972(0.020)	7.636(1.636)
	B-direct	0.994(0.006)	9.706(2.076)	1.000(0.000)	15.027(3.010)
Local individual-level	O	0.861(0.057)	4.722(1.067)	0.974(0.025)	7.556(1.725)
	B-direct	0.991(0.009)	9.229(1.877)	1.000(0.000)	14.700(2.893)

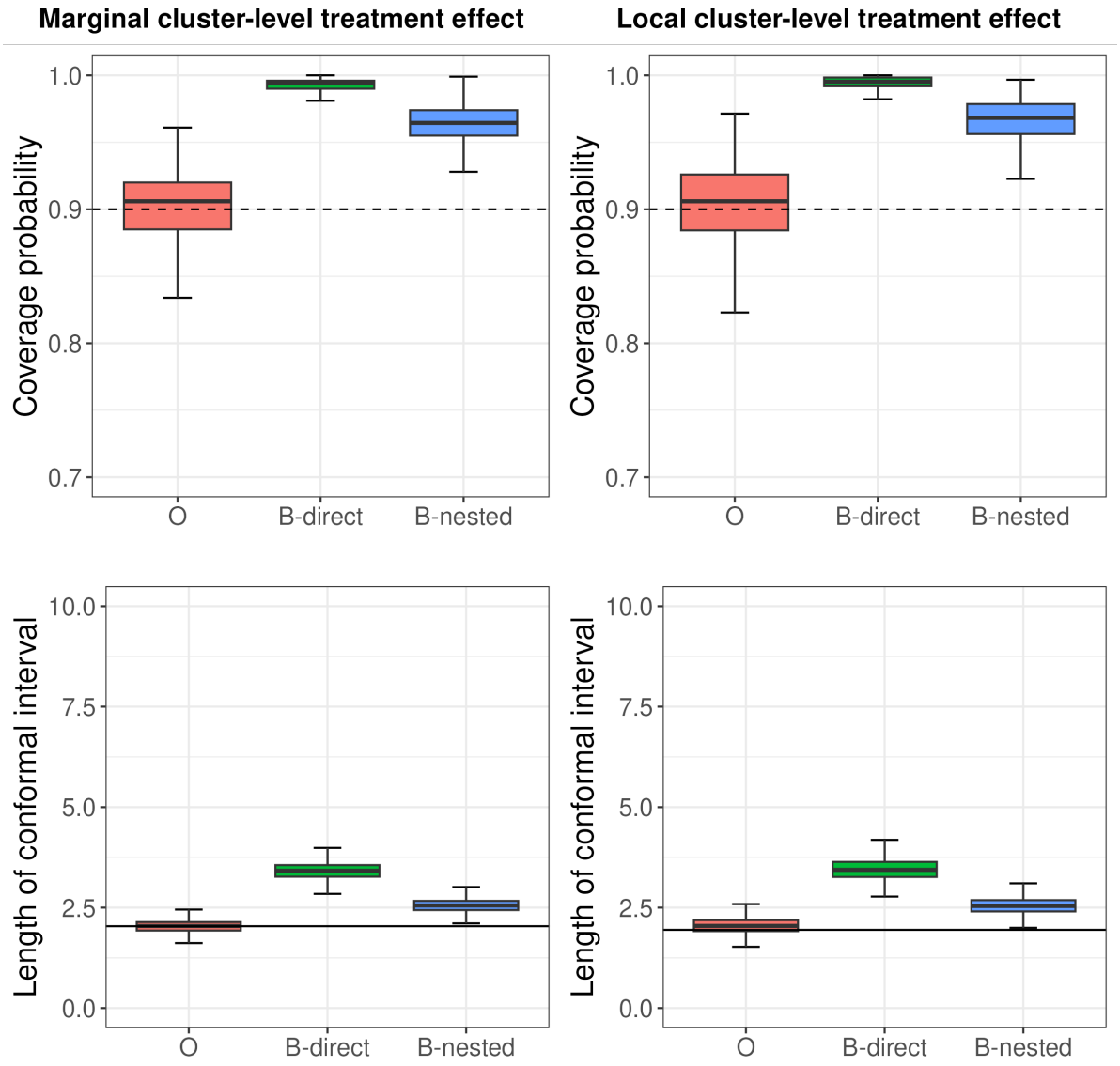

Figure 1: Simulation results (boxplot) for the marginal (left column) and local (right column, conditioning on $\{R_{i1} \geq 2, R_{i2} = 1\}$) cluster-level treatment effects with $m = 500$. In the upper panels, the dashed line is the target 90% coverage probability. In the lower panels, the solid line 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.

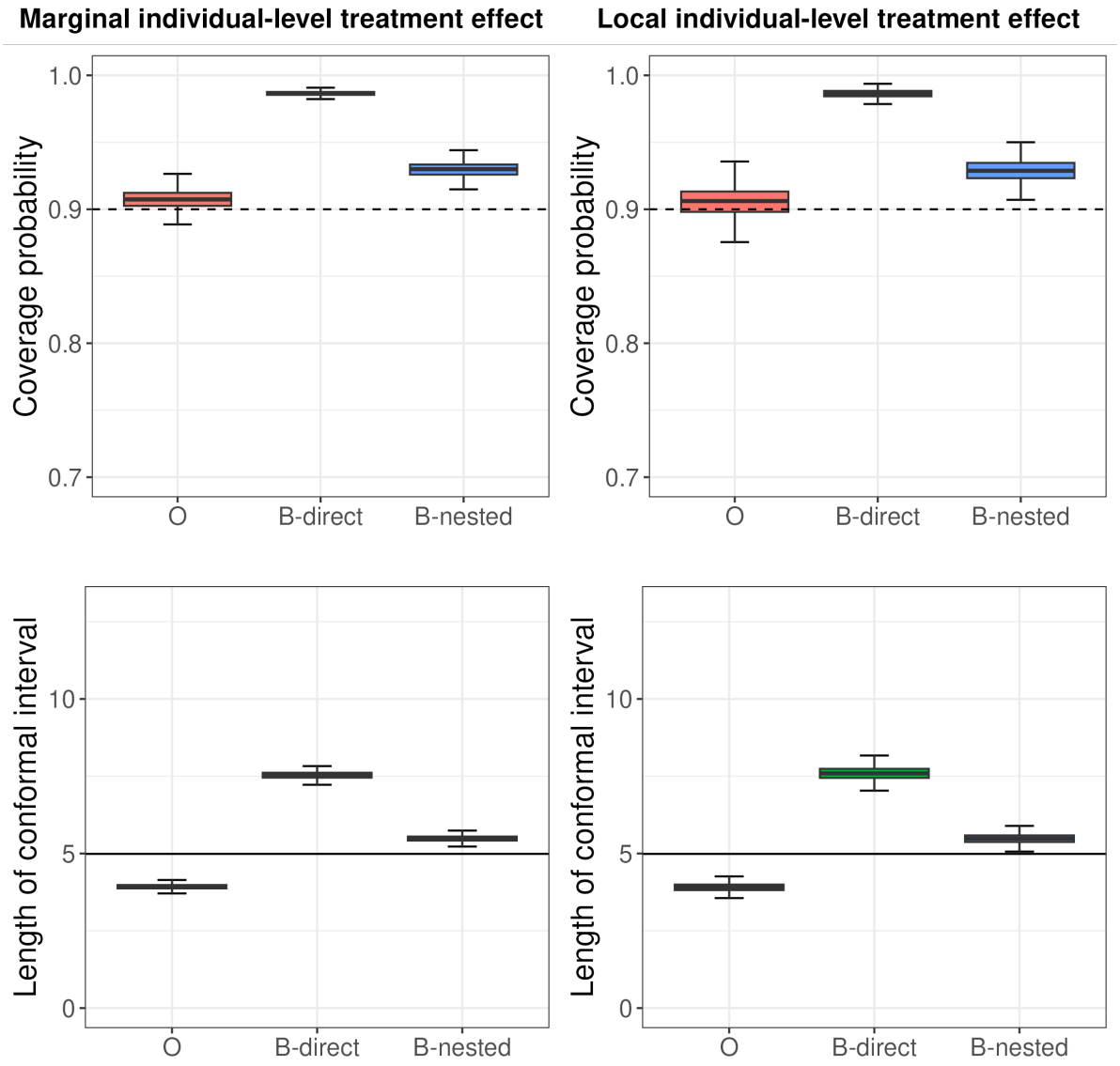

Figure 2: Simulation results (boxplot) for the marginal (left column) and local (right column, conditioning on $\{|X_{ij1}| < 0.5\}$) individual-level treatment effects with $m = 500$. In the upper panels, the dashed line is the target 90% coverage probability. In the lower panels, the solid line is 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.

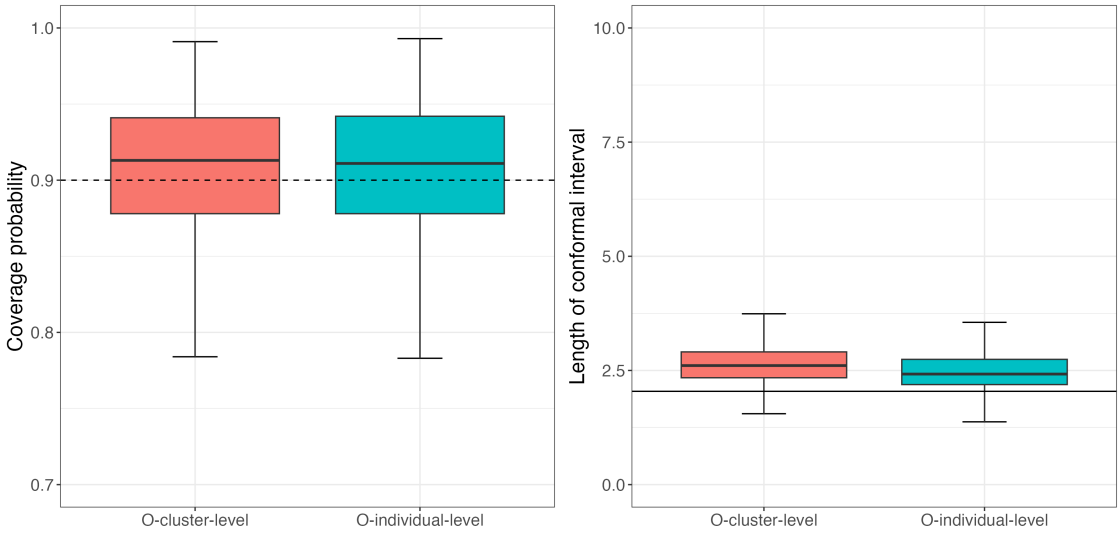

Figure 3: Comparison of prediction models based on cluster-level data ("O-cluster-level") versus individual-level data ("O-individual-level") in terms of coverage probability (left panel) and length of conformal intervals (right panel).

D.4 Simulation for cluster treatment effect based on average of individual-level prediction models

As mentioned in Section 3.1 of the main paper, the prediction model in Algorihtm 1 can be replaced by the average of individual-level predictions. Here, we compare this approach with our original proposal based on cluster-level data in our simulations, and Figure 3 shows that this new approach can slightly narrow the length of conformal intervals on average while maintaining validity.

D.5 Simulation for cluster treatment effect conditioning on cluster-averaged covariates

To assess the empirical impact of different choices of covariate subgroups, we have replicated the cluster-level simulation using the subgroup $\Omega_C = \{\overline{X}_{i1} \leq 0.7\}$. We compared two prediction models: one that includes all cluster-level covariates $\overline{B}_i$ ("O-X1-included") and another that excludes $\overline{X}_{i1}$ ("O-X1-excluded"). The results, reported in Figure 4, show that

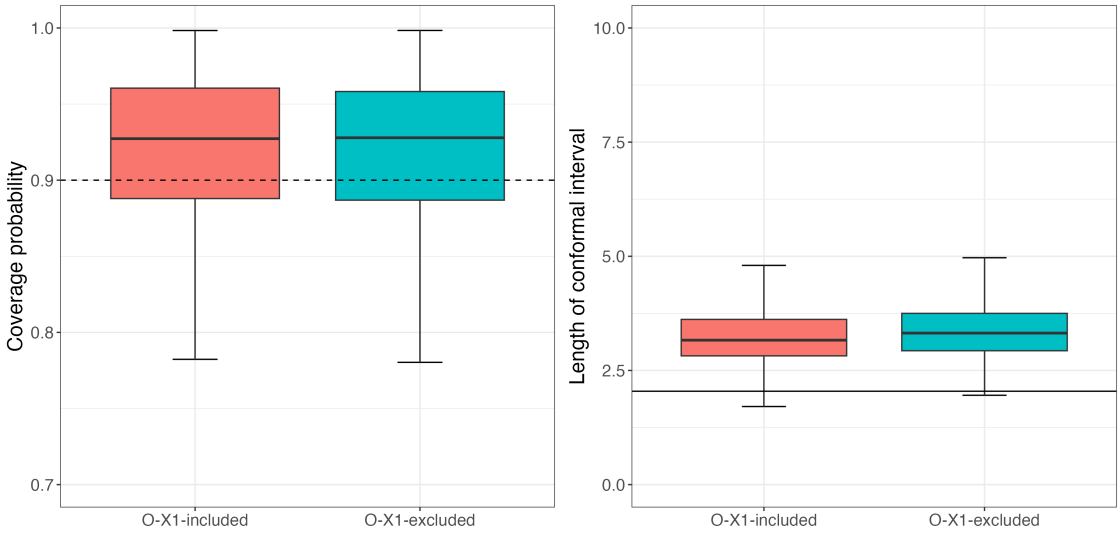

Figure 4: Comparison of prediction models for the cluster-level treatment effect in subgroup $\overline{X}_{i1} \leq 0.7$ with $\overline{X}_{i1}$ included (left panel) or excluded (right panel) in the prediction model.

the proposed conformal intervals remain valid under this alternative subgroup definition, with intervals tending to be slightly more conservative. As expected, omitting $\overline{X}_{i1}$ from the prediction model yields modestly wider intervals, reflecting reduced predictive accuracy, but the coverage property remains intact.

D.6 Data application for local treatment effects

In addition to the marginal analysis in the main paper, we performed subgroup analyses on individuals with severe baseline pain (baseline PEGS score equal to or larger than 7, 40% participants) and moderate baseline pain (baseline PEGS score between 4 and 7, 50% participants). We summarize the results in Table 3, which shows no difference in the length of intervals or the fraction of negatives between the two subgroups. Overall, these analyses indicate treatment benefits for a small subset of the clusters/individuals, which drive the overall treatment effect signal provided in previous analyses (DeBar et al., 2022; Wang et al., 2024).

Table 3: Summary results of data application for individual-level treatment effects on covariate subgroups. For both the length of intervals and the fraction of negatives, we present the average and standard error over 100 runs.

Coverage probability	Subgroup: severe baseline pain		Subgroup: moderate baseline pain	
	Length of intervals	Fraction of negatives	Length of intervals	Fraction of negatives
90%	6.538(0.571)	0.072(0.042)	6.508(0.538)	0.076(0.042)
80%	4.982(0.465)	0.120(0.051)	4.928(0.417)	0.132(0.053)
70%	3.935(0.353)	0.180(0.064)	3.960(0.349)	0.185(0.068)
60%	3.180(0.262)	0.243(0.067)	3.221(0.296)	0.241(0.079)

E Extensions to cluster observational studies

We focus on clustered observational studies with a cluster-level treatment assignment, and extend our development to accommodate a non-randomized treatment assignment mechanism. In clustered observational studies, just like the CRT setting, individual observations within a cluster are typically correlated. Denoting $[m]$ as the set $\{1, \ldots, m\}$, the "strong ignorability" condition in this context can be expressed as follows: for any cluster $i \in [m]$, the set $\{(Y_{i,1}(0), Y_{i,1}(1)), \cdots, (Y_{i,N_i}(0), Y_{i,N_i}(1))\}$ is independent of A_i given $\mathbf{B}_i := (X_{i,1}, \cdots, X_{i,N_i}, R_i)$. Here, R_i denotes the cluster-level covariates and $X_{i,j}, j \in [N_i]$ represent individual-level covariates.)

The joint distribution of $\mathbf{Y(0)}, \mathbf{B}$ among the treated, is $P_{\mathbf{Y(0)},\mathbf{B}|A=1} = P_{\mathbf{Y(0)}|\mathbf{B}} \cdot P_{\mathbf{B}|A=1}$, where $\mathbf{Y(0)} := (Y_{\bullet,1}(0), \cdots, Y_{\bullet,N_\bullet}(0))$, and $\mathbf{B} := (X_{\bullet,1}, \cdots, X_{\bullet,N_\bullet}, R_\bullet)$. For the data $\mathbf{Y(0)}$ already observed in the current study, the data are collected under the distribution $P_{\mathbf{Y(0)},\mathbf{B}|A=0} = P_{\mathbf{Y(0)}|\mathbf{B}} \cdot P_{\mathbf{B}|A=0}$. Consequently, there is a covariate shift in the target distribution.

We next describe our inferential targets. Suppose we have newly observed covariates $\mathbf{B}_{test}$ in the treatment group ($A = 1$). At the cluster level, with a pre-specified level $1 - \alpha$, our objective is to construct a prediction interval $\widetilde{C}(\overline{B}_{test})$ that contains the unobserved potential outcome $\overline{Y}_{test}(0)$ with probability $1 - \alpha$, i.e.,

$$P\{\overline{Y}_{test}(0) \in \widetilde{C}(\overline{B}_{test}) | A = 1\} \geq 1 - \alpha.$$

If $\overline{Y}(1)$ is directly observed, the prediction interval for the treatment effect is then $\overline{Y}(1) -$

$\widetilde{C}(\overline{B}_{test})$. On the other hand, if $\overline{Y}(1)$ is not observed, we can still construct a prediction interval for $\overline{Y}(1)-\overline{Y}(0)$ by leveraging the fact that $\overline{Y}(1)$ is sampled from the same distribution as the target distribution (i.e., there is no covariate shift). We apply the method described in the main text to construct a prediction interval $\widetilde{C}_{C,1}(\overline{B}_{test})$ with level $1-\alpha$ for $\overline{Y}(1)$. We then combine this prediction interval with the one for $\overline{Y}(0)$ (also at level $1-\alpha$) to achieve a coverage guarantee of $1-2\alpha$ for $\overline{Y}(1)-\overline{Y}(0)$, following a similar approach to that outlined in equation (6) of the main text.

Similarly, at the individual level, we aim to construct a prediction interval $\widetilde{C}(B_{test})$ for $Y(0)$, the unobserved individual-level potential outcome, with coverage probability $1-\alpha$, i.e.,

$$P(Y(0) \in \widetilde{C}(B_{test})|A=1) \geq 1-\alpha.$$

If $Y(0)$ is directly observed, the prediction interval for the treatment effect is given by $Y(1)-\widetilde{C}(B_{test})$. If $Y(1)$ is not directly observed, the method for constructing a prediction interval with level $1-2\alpha$ is nearly identical to the approach used at the cluster level. The key difference is that we construct an individual prediction interval $\widetilde{C}_{I,1}(B_{\text{test}})$ with level $1-\alpha$ for $Y(1)$ using the approach described in the main text.

Next, we use the technique mentioned in Tibshirani et al. (2019) to construct prediction intervals for $\overline{Y}(0)$ and $Y(0)$, to achieve the aforementioned inferential targets.

- At the cluster level, since each cluster is sampled i.i.d. from the cluster super-population, the technique for constructing the prediction for $\overline{Y}(0)$ is almost identical with the case studied in Tibshirani et al. (2019), except that we replace the individual-level covariates in Tibshirani et al. (2019) with covariates from the whole cluster $\mathbf{B}$ in constructing the weight function (described below). Hence, when covariate shift exists, we can directly follow the method proposed in Tibshirani et al. (2019) by constructing the prediction set as

$$\hat{C}_m(\overline{B}) = \left\{ y \in \mathbb{R} : S(\overline{B}, y) \leq Q_{1-\alpha}\Big(\sum_{i=1}^{m} p_i^w(\mathbf{B}_i)\delta_{S(\overline{B}_i, y_i)} + p_{m+1}^w(\mathbf{B})\delta_{+\infty} \Big) \right\}.$$

Here we let

$$p_i^w(\mathbf{B}_i) = \frac{w(\mathbf{B}_i)}{\sum_{j=1}^m w(\mathbf{B}_j) + w(\mathbf{B})}, \quad \text{and} \quad p_{m+1}^w = \frac{w(\mathbf{B})}{\sum_{j=1}^m w(\mathbf{B}_j) + w(\mathbf{B})},$$

where the weight function $w(\mathbf{B}) := dP_{\mathbf{B}|A=1}(\mathbf{B})/dP_{\mathbf{B}|A=0}(\mathbf{B})$. Note that we utilize $\mathbf{B}$ in the weight function instead of $\overline{B}$ because the aforementioned strong ignorability assumption is conditioned on $\mathbf{B}$ rather than on $\overline{B}$. Additionally, $S(\overline{B}_i, y_i), i \in [m+1]$, is the non-conformity score of the i-th cluster with $A = 0$. Notice that here there is a slight abuse of notation as we let y_i be the response that we are interested in. For example, in the above, y_i can be $\overline{Y}_i(0)$. Additionally, without loss of generality, we also assume the non-conformity score function $S(\cdot, \cdot)$ is pre-trained via sample splitting.

- At the individual level, constructing prediction sets becomes more challenging due to the violation of the original exchangeability assumption under this clustered observational data structure. To address this issue, we extend the generalized conformal prediction method from Tibshirani et al. (2019) to hierarchically structured data. In this discussion, we focus on the case where all clusters have equal cluster sizes, denoted by $M > 0$. We will address the case of clusters with random sizes in future work.

  Suppose that the first n clusters are involved in the trial data with treatment $A = 0$ and the $(n+1)$-th one is a new cluster with treatment $A = 1$. We denote $C_i = (s_{i1}, \cdots, s_{iM}), i \in [n+1]$ as a collection of non-conformity scores (associated with $Y(0)$ and B) in the i-th cluster and $g(s_{i1}, \cdots s_{iM})$ as the joint density function of $s_{i1}, \cdots s_{iM}|A = 0$. Then, we can write the joint distribution of non-conformity scores among these $n+1$ clusters to be

$$f(C_1, \cdots C_{n+1}) = \Pi_{i=1}^{n+1} h_i(\mathbf{B}_i) g(s_{i1}, \cdots, s_{iM}),$$

  where $h_i(\mathbf{B}_i) = 1$ for $i \in [n]$ and $h_{n+1}(\mathbf{B}_{n+1}) = w(\mathbf{B}_{n+1}) = dP_{\mathbf{B}|A=1}(\mathbf{B}_{n+1})/dP_{\mathbf{B}|A=0}(\mathbf{B}_{n+1})$. Here, because clusters are independent of each other, we are able to decompose $f(C_1, \cdots C_{n+1})$

into products of cluster-specific densities. Since the first n clusters are collected under $A_i = 0$, no reweighting is needed, which makes $h_i(\mathbf{B}_i) = 1$ for $i \in [n]$. However, since the $(n+1)$-th cluster is observed under $A_i = 1$, there is a covariate shift and we need to apply propensity score weighting.

Next, under the within-cluster exchangeability condition, we further have

$$g(s_{i\sigma(1)}, \cdots, s_{i\sigma(M)}) = g(s_{i1}, \cdots, s_{iM}), \quad \text{for all} \quad i \in [n+1],$$

where $\sigma(1), \ldots, \sigma(M)$ is any permutation of indices $1, \ldots, M$. Let E_z denote the observed event that

$$\{\{S_{1,1}, \cdots, S_{1,M}\}, \cdots, \{S_{n+1,1}, \cdots, S_{n+1,M}\}\} = \{\{s_{1,1}, \cdots, s_{1,M}\}, \cdots, \{s_{n+1,1}, \cdots, s_{n+1,M}\}\}.$$

Since the dataset can be labeled in any order, this equality only assume these two sets are the same, regardless of the order. We further define $\{\sigma(1,1), \ldots, \sigma(1,M), \ldots, \sigma(n+1,1), \ldots, \sigma(n+1,M)\}$ as a permutation of indices $\{(1,1), \ldots, (1,M), \ldots, (n+1,1), \ldots, (n+1,M)\}$ by first permuting the cluster indices (the first entry) and then the individual indices (the second entry). Following the proof procedure of Lemma 3 in Tibshirani et al. (2019) but in a hierarchical exchangeable setting, we obtain

$$\begin{aligned} P(S_{n+1,1} = s_{i,j} | E_z) &= \frac{\sum_{\sigma:\sigma(n+1,1)=(i,j)} f(\{s_{\sigma(1,1)}, \cdots, s_{\sigma(1,M)}\}, \cdots, \{s_{\sigma(n+1,1)}, \cdots, s_{\sigma(n+1,M)}\})}{\sum_{\sigma} f(\{s_{\sigma(1,1)}, \cdots, s_{\sigma(1,M)}\}, \cdots, \{s_{\sigma(n+1,1)}, \cdots, s_{\sigma(n+1,M)}\})} \\ &= \frac{\sum_{\sigma:\sigma(n+1,1)=(i,j)} \Pi_{i=1}^{n+1} h_i(\mathbf{B}_{\sigma(i)}) g(\{s_{\sigma(i,1)}, \cdots, s_{\sigma(i,M)}\})}{\sum_{\sigma} \Pi_{i=1}^{n+1} h_i(\mathbf{B}_i) g(\{s_{\sigma(i,1)}, \cdots, s_{\sigma(i,M)}\})} \\ &= \frac{w(B_i)}{M \cdot w(B_1) + \cdots + M \cdot w(B_n) + M \cdot w(B_{n+1})} := p_{i,j}^w. \end{aligned}$$

The second equality follows from the fact that $h_i(\cdot) = 1$ for all $i \in [n]$ and the factors involving g are all cancelled. Therefore, the final prediction set with coverage level

$1-\alpha$ is constructed as

$$\widetilde{C}_n(B_{n+1,1}) = \Big\{ y \in R : S(B_{n+1,1}) \leq Q_{1-\alpha}\Big(\sum_{i=1}^{n} \sum_{j=1}^{M} p_{i,j}^{w} \delta_{S(B_{i,j},y_{i,j})} + \sum_{j=1}^{M} p_{n+1,j}^{w} \delta_{+\infty} \Big) \Big\}.$$

References

DeBar, L., Mayhew, M., Benes, L., Bonifay, A., Deyo, R. A., Elder, C. R., Keefe, F. J., Leo, M. C., McMullen, C., Owen-Smith, A., 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.

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

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

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

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

Yang, S., Moerbeek, M., Taljaard, M., and Li, F. (2023). Power analysis for cluster randomized trials with continuous coprimary endpoints. *Biometrics*, 79(2):1293–1305.