Bridging Fairness and Efficiency in Conformal Inference: A Surrogate-Assisted Group-Clustered Approach

Chenyin Gao [1] **Peter B. Gilbert** [2] **Larry Han** [2 3]

Abstract

Standard conformal prediction ensures marginal coverage but consistently undercovers underrepresented groups, limiting its reliability for fair uncertainty quantification. Group fairness requires prediction sets to achieve a user-specified coverage level within each protected group. While group-wise conformal inference meets this requirement, it often produces excessively wide prediction sets due to limited sample sizes in underrepresented groups, highlighting a fundamental tradeoff between fairness and efficiency. To bridge this gap, we introduce Surrogate-Assisted Group-Clustered Conformal Inference (SAGCCI), a framework that improves efficiency through two key innovations: (1) clustering protected groups with similar conformal score distributions to enhance precision while maintaining fairness, and (2) deriving an efficient influence function that optimally integrates surrogate outcomes to construct tighter prediction sets. Theoretically, SAGCCI guarantees approximate group-conditional coverage in a doubly robust manner under mild convergence conditions, enabling flexible nuisance model estimation. Empirically, through simulations and an analysis of the phase 3 Moderna COVE COVID-19 vaccine trial, we demonstrate that SAGCCI outperforms existing methods, producing narrower prediction sets while maintaining valid group-conditional coverage, effectively balancing fairness and efficiency in uncertainty quantification.

[1]Harvard University, Department of Biostatistics, Boston, MA, USA [2]Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutch Cancer Center, Seattle, WA, USA [3]Northeastern University, Department of Public Health and Health Sciences, Boston, MA, USA. Correspondence to: Larry Han <lar.han@northeastern.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Personalized medicine calls for precise individualized predictions to guide patient-level decision-making. Equally critical is robust uncertainty quantification for these predictions, particularly in contexts like vaccine safety or efficacy studies, where personalized vaccine recommendations can influence critical individual-level decisions. Examples include tailoring guidance for vulnerable populations, such as immunocompromised individuals, or for individuals evaluated based on antibody levels against specific pathogens.

An ideal prediction set for uncertainty quantification must meet two essential criteria: (1) valid coverage in finite samples without relying on stringent distributional assumptions, and (2) sufficiently short prediction set lengths to ensure practical utility. Conformal prediction is a powerful method for uncertainty quantification, offering distribution-free coverage guarantees alongside reasonable prediction set lengths (Vovk et al., 2009; Lei et al., 2013). However, standard conformal prediction methods ensure marginal coverage only, i.e., that prediction sets contain the true labels on average across the population. While effective in providing distribution-free guarantees, these methods often fail to provide adequate coverage for sub-groups, which may result in disparities and lead to ethical or legal implications. Fair coverage is crucial in socially consequential domains, where prediction intervals can help guide decision-making related to access to resources, opportunities, or fair treatment. For example, if a mortgage lender's prediction intervals systematically under-cover minority borrowers, they may face inflated interest rates or higher denial rates. In healthcare, if limited supplies of critical antibiotics demand triaging decisions, under-coverage for certain demographics could result in denying them lifesaving medications.

Efforts to design conformal inference for fair predictions, such as the approach proposed by Vovk (2012), recommend performing conformal inference separately for each subgroup to ensure group-conditional coverage. However, this strategy often produces excessively wide prediction intervals due to limited sample sizes within smaller sub-groups, reducing their practical value. There is a growing need to develop methods that design more efficient prediction sets for sub-groups, particularly in contexts where fairness and

precision are critical. Two promising avenues to address this challenge are clustering and the use of surrogate outcomes. Clustered conformal inference, introduced by Ding et al. (2024) for outcome-dependent inference, provides a more efficient solution for achieving group-conditional coverage. Instead of performing group-wise conformal inference, this method clusters groups with similar conformalized score distributions into larger clusters, allowing for more efficient individualized predictions while maintaining fairness. Additionally, surrogate outcomes – such as biological markers or machine-learned noisy predictions – offer another powerful approach to improving efficiency in uncertainty quantification. These surrogates, which are often easier to observe or collect than the primary outcomes of interest, enhance efficiency, especially in scenarios where primary outcomes are challenging to measure or largely unavailable. For instance, in vaccine studies, immune correlates like neutralizing antibody titers have been shown to serve as effective surrogates for predicting vaccine efficacy (Gilbert et al., 2022; 2024).

Our contributions In this paper, we propose a novel Surrogate-Assisted Group-Clustered Conformal Inference (SAGCCI) framework that achieves group-conditional coverage while improving the efficiency of uncertainty quantification through two key innovations:

1. Clustering: By pooling groups with similar score distributions, we leverage clustering algorithms to deal with sub-groups with small sample sizes.
2. Surrogate-assisted efficient influence function (EIF): We develop an EIF-based approach for constructing conformalized prediction sets, effectively incorporating surrogate outcomes to maximize efficiency.

Theoretically, we demonstrate that our method guarantees approximate group-conditional coverage, provided the clustering algorithm pools groups accurately with minimal errors and the model estimation satisfies standard convergence rate conditions, which are common in flexible machine learning models (Chernozhukov et al., 2018). We validate the practical effectiveness of SAGCCI through extensive synthetic simulations and a real-data analysis of the phase 3 Moderna COVE COVID-19 vaccine, showcasing its ability to improve uncertainty quantification for individualized predictions. Our proposal advances statistical theory and machine learning applications in personalized medicine, enabling fairer and more efficient uncertainty quantification of predictions. Our code is publicly available at `https://github.com/Gaochenyin/SurrConformalDR`.

2. Related work in surrogates

The concept of statistical surrogacy was introduced by Prentice (1989), who defined criteria for surrogate evaluation. Since then, numerous surrogate validation methods have been developed to assess surrogate outcomes, which Conlon et al. (2017) broadly categorized into two main approaches: causal effects (Gilbert & Hudgens, 2008) and causal association frameworks (Li et al., 2010; Alonso et al., 2016). For a detailed review of these methods and recent advancements, see Elliott (2023). Surrogate validation aims to identify outcomes that can replace primary endpoints in clinical trials, thereby facilitating the expedited approval of treatments, particularly for critical illnesses. Regulatory agencies like the FDA have relied on surrogate endpoints to accelerate drug approvals (FDA, 1992; 2021). However, surrogate outcomes are not always reliable, as illustrated by cases where treatments approved based on surrogate markers, such as arrhythmia suppression, ultimately led to adverse outcomes like increased mortality in follow-up studies (Fleming & DeMets, 1996). To address these challenges, surrogate validation methods often impose stringent assumptions, such as the strong statistical surrogate condition (Prentice, 1989), which can be difficult to verify in practice. For example, Athey et al. (2019) relied on such assumptions to integrate experimental and observational datasets.

In contrast, our work treats surrogates as auxiliary information or "helper covariates" (Xia & Wainwright, 2024) rather than as substitutes for primary outcomes. This perspective aligns with recent work that leverages surrogate outcomes to improve the efficiency of estimating long-term or primary treatment effects. For instance, Athey et al. (2020) proposed combining experimental data (containing only surrogates) with observational data (containing both surrogates and primary outcomes) to estimate long-term treatment effects. Chen & Ritzwoller (2023) considered this framework and derived semiparametric efficiency bounds. Other studies, including Cheng et al. (2021) and Kallus & Mao (2024), have explored efficient estimation methods that rely on a limited number of primary outcomes alongside abundant surrogates, without requiring strong statistical surrogacy. Furthermore, Imbens et al. (2024) examined sequentially structured surrogates and addressed challenges in identifying average treatment effects when unmeasured confounders are present. Gao et al. (2024) leveraged surrogates to improve the efficiency of prediction sets for individual treatment effects, showing improved efficiency compared to existing methods (Lei & Candès, 2021; Yang et al., 2024).

3. Preliminaries

3.1. Notations

Let $\mathcal{X} \subseteq R^{d_X}$, $\mathcal{Z} = \{1, \cdots, M\}$, and $\mathcal{Y}$ denote the space of features, protected groups (or sensitive attributes, e.g., race/gender), and the primary outcomes, respectively. Define X, Z, and Y as random variables taking values in $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{Y}$, respectively. Additionally, suppose we have access

to surrogate outcomes $S \in \mathcal{S} \subseteq \mathbb{R}^{d_S}$ for each individual. These surrogates, such as biomarkers (e.g., neutralizing antibody titers), may serve as predictors of the outcome Y.

We distinguish between two types of data: smaller source data of size n_{D1}, where the primary outcome Y is observed, and larger target data of size n_{D0}, where Y is missing. This setup is motivated by real-world situations where labeled source data may be costly or require lengthy follow-up periods, while large amounts of unlabeled target data can be cheaply obtained from existing databases such as electronic health records (Cheng et al., 2021). Let $D \in \{0, 1\}$ indicate the data origin, with $D = 1$ representing source data and $D = 0$ representing target data. In summary, we observe a source dataset $\{(X_i, Z_i, S_i, Y_i, D_i = 1)\}_{i=1}^{n_{D1}}$ and a target dataset $\{(X_i, Z_i, S_i, Y_i = \text{NA}, D_i = 0)\}_{i=n_{D1}+1}^{N}$, where NA means "not available" (missing), and $N = n_{D1} + n_{D0}$ is the total sample size.

In our framework, we treat the baseline covariates X and the surrogate outcomes S separately, as the distributional differences in surrogates across the source and target datasets may be fully explained by the covariates (e.g., Assumption 4.2(a)). This enables us to leverage surrogates from both datasets to improve the estimation of primary outcomes, conditional on the covariates (Theorem 4.4).

3.2. Conformalized prediction

Let $W = (X, S) \in \mathcal{W} = \mathcal{X} \times \mathcal{S}$ be a set of predictors, and let $R(W, Y) : \mathcal{W} \times \mathcal{Y} \to \mathbb{R}$ denote the non-conformity score function (Vovk et al., 2005). For example, common choices of $R(W, Y)$ include the regression residual $R(W, Y) = |Y - \widehat{E}(Y \mid W)|$ (Lei et al., 2018) or the conformalized quantile residual $R(W, Y) = \max\{\widehat{q}_{\alpha/2}(W) - Y, Y - \widehat{q}_{1-\alpha/2}(W)\}$, where $\widehat{q}_{\alpha/2}(\cdot)$ and $\widehat{q}_{1-\alpha/2}(\cdot)$ are estimated conditional quantiles (Romano et al., 2019). Given independent and identically distributed data $V_i = (W_i, Z_i, Y_i) \sim P_{D_1}$ drawn from the source distribution P_{D_1} for $i = 1, \ldots, n_{D1}$, the prediction sets can be constructed based on the non-conformity scores $R(W, Y)$ such that $C(W) = C(W; r_\alpha) = \{y : R(W, y) \le r_\alpha\}$, given a threshold r_α for a user-specified miscoverage level $\alpha \in (0, 1)$.

The threshold r_α should be chosen such that, for any data $V = (W, Z, Y) \sim P_{D_0}$ drawn from the target distribution, class label Y has at least $1-\alpha$ probability of being included in the prediction set $C(W; r_\alpha)$, that is,

$$P_{V \sim P_{D_0}}(Y \in C(W; r_\alpha)) \ge 1 - \alpha, \tag{1}$$

where the probability is taken over the marginal distribution of the source data and the future observation from the target data V. The marginal coverage guarantee (1) does not imply covariate-conditional coverage, $P_{V \sim P_{D_0}}(y \in C(W; r_\alpha^z) \mid W = w) \ge 1 - \alpha$, which is more desirable in practice. However, exact covariate-conditional coverage is known to be generally unattainable, as shown in Foygel Barber et al. (2021). Some efforts have been made to achieve approximate covariate-conditional coverage by designing better scores (Romano et al., 2019; 2020) or by modifying the conformal procedure (Guan, 2023; Gibbs et al., 2023).

3.3. Fairness notion in coverage

Despite the impossibility of achieving exact covariate-conditional coverage, weaker coverage guarantees can be obtained by conditioning on a set of W rather than limiting W to a specific value w. This approach, referred to as group-conditional coverage, aligns closely with the concept of group fairness (Dwork et al., 2012), or demographic parity (Kusner et al., 2017). Group fairness requires certain statistical metrics to be equalized across different levels of a protected attribute, such as race or gender (Makhlouf et al., 2021). For further discussion on related group fairness notions, including equalized odds and equal opportunity, see Hardt et al. (2016).

Next, we claim that the group-specific prediction set $C(W; r_\alpha^z)$ satisfies group fairness, such that

$$P_{V \sim P_{D_0}}(y \in C(W; r_\alpha^z) \mid Z = z) \ge 1 - \alpha, \tag{2}$$

where r_α^z is the threshold for group z. Therefore, the prediction sets defined in (2) are fair across all protected groups in terms of their coverage, irrespective of group size.

Remark 3.1. If a given prediction model achieves fair coverage, it does not necessarily imply fairness in point predictions, especially when the lengths of the prediction sets are heterogeneous. Therefore, fair coverage can be viewed as a fairness metric for interval predictions, as it accounts for the uncertainty in point predictions.

To achieve the desired coverage across all levels of the protected variable as defined in (2), it is natural to estimate the thresholds r_α^z separately for each protected group (Vovk, 2012), and further account for the covariate shifts across the protected groups as in Gibbs et al. (2023). However, this group-wise strategy can be overly conservative when some protected groups are small in size, leading to prediction sets that are excessively large and of limited practical use in finite samples. Conversely, if the threshold r_α is estimated using the entire dataset without splitting by group, the standard conformal inference procedure cannot guarantee group-conditional coverage. In such cases, certain groups may experience substantial under-coverage, while others may be significantly over-covered.

4. Surrogate-assisted group-clustered conformal inference (SAGCCI)

To overcome the limitations of standard conformal inference and group-conditional conformal inference and to bridge fairness and efficiency, we propose SAGCCI, which leverages (i) clustering of protected groups based on non-conformity scores to overcome the issue of limited sample sizes in some groups and (ii) an EIF-based estimator for constructing the conformalized prediction sets, which uses surrogates to boost the efficiency of the prediction sets.

4.1. Assumptions

We require the following assumptions to ensure the identification of the threshold for the non-conformity scores, which are an essential ingredient in the construction of our EIF-based estimator.

Assumption 4.1. There exists some constant $0 < c_0 < 1/2$ such that $c_0 \leq P(D = 1 \mid x) \leq 1 - c_0$ for any x such that $f(x) > 0$.

Assumption 4.1 states that each individual should have a probability of at least c_0 of being included in the target data, conditional on any covariates with a positive likelihood of occurrence. This overlap assumption is common in the causal inference and missing data literature (Imbens & Rubin, 2015).

Assumption 4.2. (a) $D \perp S \mid X$, and (b) $D \perp Y \mid S, X$.

Assumption 4.2(a) states that the data origin indicator D is conditionally independent of the surrogates S given the observed covariates X. This implies that surrogates from the combined dataset (if available) can be utilized to improve efficiency.

Furthermore, Assumption 4.2(b) states that the data origin indicator D is conditionally independent of the primary outcome Y given the observed covariates X and the surrogates S. Since D effectively indicates whether Y is missing, Assumption 4.2(b) corresponds to the missing at random (MAR) assumption in the missing data literature (Little & Rubin, 2019). This implies that any missingness in the primary outcome is fully explained by the observed covariates and surrogates. As a result, covariates and surrogates from the source data can be used to infer information about the missing primary outcomes in the target data.

4.2. Clustered conformal prediction

In this section, we introduce a clustering approach for protected groups based on the similarity of their conformal score distributions. By grouping smaller protected groups into larger clusters, this method improves the efficiency of prediction sets, particularly for groups with limited data. This strategy builds on similar ideas from Ding et al. (2024), who used clustering to improve outcome-dependent coverage in settings with categorical outcomes across many classes.

First, we embed the empirical distribution for the group-specific scores using a vector constituted by their quantiles evaluated at discrete levels $\tau = \{0.5, \cdots, 0.9\} \cup \{1 - \alpha\}$. Various embedding methods, such as spectral embeddings or other non-linear embeddings, can be employed, where a greater distance between any two embeddings indicates a larger disparity in their score distributions.

Once these embeddings are constructed, we can efficiently implement any clustering algorithm, such as k-means, to produce the desired clusters; other cluster mappings, such as fuzzy k-means or overlapping k-means, can be applied, to obtain overlapping clusters. For a pre-specified number of clusters K, we apply the clustering algorithm to partition the embeddings accordingly. Let $\mathcal{I}^z = \{i : Z_i = z, i = 1, \ldots, N\}$ represent the indices of subjects in the protected group z. However, for subjects with unknown protected groups, or for protected groups with extremely small sizes (e.g., $|\mathcal{I}^z| < (1/\alpha) - 1$), the empirical $(1 - \alpha)$-th quantile of the scores for these groups does not exist. To address this issue, these groups are assigned to a null cluster. We define the cluster mapping from protected groups to cluster labels as $\widehat{h}$, where $\widehat{h} : \mathcal{Z} \to \{1, \ldots, K\} \cup \{\text{null}\}$.

If the score distributions of groups assigned to the same cluster by $\widehat{h}$ are sufficiently similar, approximate group-conditional coverage guarantees can be established. Similarity of the scores is measured using the Kolmogorov-Smirnov (KS) distance, which quantifies the maximum discrepancy between the cumulative distribution functions of two random variables X and Y: $\text{KS}(X, Y) = \sup_{\lambda \in \mathbb{R}} |P(X \leq \lambda) - P(Y \leq \lambda)|$.

Lemma 4.3. *Let $R^z = \{R(W_i, Y_i) : Z_i = z\}$ be the non-conformity scores for group z. Suppose the clustering map $\widehat{h}$ satisfies KS$(R^z, R^{z'}) \leq \epsilon$ for every pair (z, z') such that $\widehat{h}(z) = \widehat{h}(z') = k$. Then, for any protected group z with $\widehat{h}(z) = k$, the following holds:*

$$P_{V \sim P_{D_0}}(y \in C(W; r_\alpha^k) \mid Z = z) = P_{V \sim P_{D_0}}(R(W, y) \leq r_\alpha^k \mid Z = z) \geq 1 - \alpha - \epsilon,$$

where r_α^k is the threshold for cluster k.

Lemma 4.3 demonstrates that when the score distributions of two protected groups assigned to the same cluster are sufficiently similar, as measured by their KS distance, the cluster-specific thresholds r_α^k can be used to construct prediction sets for the associated groups with guaranteed coverage. This approach enhances the precision of the prediction sets, as the thresholds are estimated using larger, pooled clusters, thereby benefiting from increased data size. The

number of clusters K should be pre-specified, and we provide a simple heuristic in the Appendix for selecting this parameter in practice.

4.3. Efficient influence function

We now derive the EIF for r_α (or r_α^k for the k-th cluster), which is a functional derivative that characterizes how sensitive the estimand is to changes in the data generation distributions P_{D_1} and P_{D_0}. The EIF, also known as the canonical gradient (Van der Laan et al., 2011), is a fundamental tool for achieving statistical efficiency. Leveraging the EIF, we construct estimators for r_α that incorporate surrogate outcomes, enabling the estimator to achieve local semiparametric efficiency.

Theorem 4.4. *Under Assumptions 4.1 and 4.2, the EIF $\psi_\alpha^r(V; r_\alpha)$ without surrogates and the EIF $\psi_\alpha(V; r_\alpha)$ with surrogates for r_α is, up to a proportionality constant,*

$$\psi_\alpha^r(V; e_D, m, r_\alpha) = (1-D)\{m(r_\alpha, X) - (1-\alpha)\} + D\pi_D(X)\{\mathbf{1}(R \le r_\alpha) - m(r_\alpha, X)\},$$

and

$$\psi_\alpha(V; e_D, m, \tilde{m}, r_\alpha) = \psi_\alpha^r(V; e_D, m, r_\alpha) + [D\pi_D(X) - \{1-e_D(X)\}]\{m(r_\alpha, X) - \tilde{m}(r_\alpha, W)\},$$

where $e_D(X) = P(D = 1 \mid X)$ is the propensity score of being the source data, $\pi_D(X) = \{1 - e_D(X)\}/e_D(X)$ is the inverse odds of observing primary outcomes, $\tilde{m}(r_\alpha, W) = P(R \le r_\alpha \mid W)$ is the conditional cumulative distribution function (CDF) of the non-conformity scores incorporating surrogates, and $m(r_\alpha, X) = P(R \le r_\alpha \mid X)$ is the conditional CDF without surrogates, both evaluated at r_α.

We refer to e_D, m, and $\tilde{m}$ as nuisance functions since they are not of our interest but are unknown and required to evaluate the EIFs for r_α. Theorem 4.4 establishes that the surrogate-assisted EIF for r_α extends the EIF without surrogates, $\psi_\alpha^r(V; e_D, m, r_\alpha)$, by incorporating an additional augmentation term. This augmentation term has a conditional expectation of zero, ensuring that it introduces no bias. The purpose is to leverage the residual information provided by the surrogates S in both datasets, thereby boosting the estimation efficiency. The efficiency gain achieved through the inclusion of surrogates is formally quantified in Corollary 4.5.

Corollary 4.5. *Under Assumptions 4.1 and 4.2, the efficiency gain of leveraging the surrogates is given by*

$$V_{\text{eff}}^r - V_{\text{eff}} = E\left[\pi_D(X)\{1-e_D(X)\}^2 \text{var}\{\tilde{m}(r_\alpha, W) \mid X\}\right],$$

where $V_{\text{eff}}^r = \text{var}\{\psi_\alpha^r(V; r_\alpha)\}$, $V_{\text{eff}} = \text{var}\{\psi_\alpha(V; r_\alpha)\}$.

Corollary 4.5 quantifies the efficiency gain achieved by incorporating surrogates in terms of the semiparametric efficiency lower bound for estimating the threshold r_α. Specifically, this efficiency gain depends on how predictive the surrogates are for the primary outcomes. The predictiveness is captured by the variance $\text{var}\{\tilde{m}(r_\alpha, W) \mid X\} = \text{var}\{P(R \le r_\alpha \mid W) \mid X\}$, which reflects the additional variability in predicting the scores R that is explained by the surrogates S beyond the covariates X. A higher variance indicates that the surrogates contribute substantial information about the primary outcomes, resulting in greater efficiency gains in estimating r_α. Moreover, the efficiency gain is influenced by the proportion of source data, as determined by the propensity scores $e_D(X)$. When $e_D(X)$ is large, the proportion of the labeled source data is high, which intuitively means there is less missingness in the primary outcomes to address. As a result, the potential efficiency gain from leveraging surrogates is reduced in such cases.

4.4. Implementation

To construct the semiparametric efficient estimator $\widehat{r}_\alpha^k$ for cluster $k = 1, \cdots, K$, we adopt the split conformal inference strategy (Lei & Wasserman, 2014). This method involves randomly spliting the combined data into two folds, $\mathcal{I}_1$ and $\mathcal{I}_2$. The first fold, $\mathcal{I}_1$, is used to train the learning algorithms (e.g., the nuisance functions e_D, m, and $\tilde{m}$). The second fold, $\mathcal{I}_2$, is used to construct prediction intervals based on the non-conformity scores.

For concreteness, we describe the split conformal inference for Y using conformalized quantile residuals (CQR) on the target data. Let $\widehat{q}_{\alpha/2}(\cdot)$ and $\widehat{q}_{1-\alpha/2}(\cdot)$ denote the quantile models for the primary outcomes Y, trained on the first fold $\mathcal{I}_1$. Additionally, the propensity score $\widehat{e}_D$ is estimated using $\mathcal{I}_1$. To mitigate the computational complexity of estimating the full conditional distributions $m(r, X)$ and $\tilde{m}(r, W)$ for infinitely many r, we adopt the localized debiased machine learning approach (Kallus et al., 2024). Under this approach, the first fold, $\mathcal{I}_1$, is further split into two sub-folds, $\mathcal{I}_{11}$ and $\mathcal{I}_{12}$. In the first sub-fold, $\mathcal{I}_{11}$, we construct an initial estimator $\widehat{r}_\alpha^{\text{init}}$ for r_α. A natural choice is the weighted split-CQR estimator from Lei & Candès (2021). In the second sub-fold, $\mathcal{I}_{12}$, the nuisance models $\widehat{m}(\widehat{r}_\alpha^{\text{init}}, X)$ and $\widehat{\tilde{m}}(\widehat{r}_\alpha^{\text{init}}, W)$ are estimated using any machine learning binary algorithm given the single initial estimator $\widehat{r}_\alpha^{\text{init}}$.

Next, we compute the non-conformity scores R for the second fold, $\mathcal{I}_2$, using the estimated quantile models $\widehat{q}_{\alpha/2}(\cdot)$ and $\widehat{q}_{1-\alpha/2}(\cdot)$. The scores are defined as $\widehat{R}(W_i, Y_i) = \max\left\{\widehat{q}_{\alpha/2}(W_i) - Y_i, Y_i - \widehat{q}_{1-\alpha/2}(W_i)\right\}$. From these scores, the score distributions for each protected group are embedded using their group-specific quantiles. We then apply k-means clustering to map the M protected groups into K clusters, learning the cluster mapping $\widehat{h}$ based

on the score embeddings. Finally, within $\mathcal{I}_2$, we identify the cluster-specific efficient estimators $\widehat{r}_\alpha^k$ for the threshold. Let $\mathcal{I}^{(k)} = \{i : \widehat{h}(Z_i) = k, i \in \mathcal{I}_2\}$ represent the indices of subjects who are assigned to cluster k by the mapping $\widehat{h}$. These estimators are defined as the smallest values satisfying the condition: $\sum_{i\in\mathcal{I}_2\cap\mathcal{I}^{(k)}} \psi_\alpha(V_i; \widehat{e}_D, \widehat{m}, \widehat{\widetilde{m}}, \widehat{r}_\alpha^k) \geq 0$, where the nuisance models $\widehat{m}$, $\widehat{\widetilde{m}}$, and $\widehat{e}_D$ are estimated from $\mathcal{I}_1$. Full implementation details are provided in Algorithm 1.

4.5. Theoretical properties

Next, we establish the theoretical properties of the estimator $\widehat{r}_\alpha^k$ for cluster k. To do so, we first list the regularity conditions for the nuisance functions:

(A1) The estimated nuisance functions $\widehat{e}_D(X)$, $\widehat{m}(r, X)$ and $\widehat{\widetilde{m}}(r, W)$ are bounded. Specifically, there exist constants π_0, m_0 and $\tilde{m}_0$ such that $|\widehat{\pi}_D(X)| \leq \pi_0$, $|\widehat{m}(r, X)| \leq m_0$, and $|\widehat{\widetilde{m}}(r, W)| \leq \tilde{m}_0$.

(A2) The estimators $\widehat{m}(r, X)$ and $\widehat{\widetilde{m}}(r, W)$ are non-decreasing with respect to r.

Using the split conformal inference strategy outlined in Section 4.4, the desired asymptotic group-conditional coverage of the prediction sets is guaranteed in Theorem 4.6.

Theorem 4.6. *Under Assumptions 4.1 and 4.2, along with regularity conditions (A1) and (A2), there exist constants C_0 and C_1 such that, for any group z satisfying $\widehat{h}(z) = k$, the following holds with probability at least $1 - \delta$ with $\delta > 0$:*

$$
\begin{aligned}
&P_{V\sim P_{D_0}}(y \in C(W; r_\alpha^k) \mid Z = z) \\
&= P_{V\sim P_{D_0}}(R(W, y) \leq r_\alpha^k \mid Z = z) \geq 1 - \alpha - \epsilon \\
&- C_0(\pi_0 + m_0 + \pi_0\tilde{m}_0)\sqrt{\frac{\log(1/\delta) + 1}{|\mathcal{I}_2 \cap \mathcal{I}^{(k)}|}} \\
&- C_1 \left\{ \|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{m}(r, X) - m(r, X)\| \right. \\
&\left. + \|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{\widetilde{m}}(r, W) - \tilde{m}(r, W)\| \right\}.
\end{aligned}
$$

Theorem 4.6 ensures a group-conditional coverage guarantee of approximately $1-\alpha$ for the cluster-specific prediction sets of Y from the target population. The slack in the coverage guarantee is the sum of three components:

1. Score distribution discrepancy: The first term, ϵ, arises due to differences in the score distributions of protected groups within the same cluster, which usually diminishes to zero for large sample sizes (e.g., the KS k-means clustering (Zhu et al., 2021)).

2. Finite sample error: The second term, obtained from empirical process theory, is proportional to $O(N^{-1/2})$, results from bounding the empirical mean of $\psi_\alpha(V; \widehat{e}_D, \widehat{m}, \widehat{\widetilde{m}}, \widehat{r}_\alpha)$ in finite samples. This error decreases as the sample size increases, given a fixed number of clusters and similar sizes for the data folds $\mathcal{I}_1$ and $\mathcal{I}_2$ used in the split inference strategy.

3. Product bias: The third term reflects the bias induced by estimation errors for the nuisance functions. The bias is negligible if either $\|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{m}(r, X) - m(r, X)\| = o(1)$ and $\|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{\widetilde{m}}(r, W) - \tilde{m}(r, W)\| = o(1)$. This property, known as rate double robustness (Chernozhukov et al., 2018), implies that small perturbations in the nuisance functions affect the coverage error only in second-order terms.

As a result, the prediction sets $C(W; r_\alpha^k)$ constructed using SAGCCI approximately achieve the desired group-conditional coverage level, up to a negligible term that vanishes with high probability, that is, $P_{V\sim P_{D_0}}(y \in C(W; r_\alpha^k) \mid Z = z) = 1 - \alpha - o(1)$ as $N \to \infty$. This property is commonly referred to as Probably Approximately Correct (PAC) coverage in the statistical literature (Krishnamoorthy & Mathew, 2009).

5. Simulation studies

We evaluate the performance of our proposed method through a series of numerical experiments. Empirical results show that SAGCCI effectively reduces disparities in group-conditional coverage while significantly shrinking prediction set sizes relative to existing methods.

5.1. Experimental setup

To simulate a challenging scenario, we assume that the source data (with observed primary outcomes) is significantly smaller than the target data. Specifically, we randomly select a small proportion, $P(D = 1) = N^{-1/10}$ of the total population as the source data ($D = 1$), with the remaining data constituting the target data ($D = 0$). The total sample sizes considered are $N = 1000, 3000, 5000, 10000$. We first generate the baseline covariates $X \in \mathbb{R}^2$ from a multivariate normal distribution, $X \sim \mathcal{N}(0, I_2)$, for the entire population.

Next, we generate the protected group variable Z with $M = 3$ levels from a multinomial distribution. Specifically, $P(Z = 1 \mid X) \propto \exp\left(-\alpha_{Z,1} - \sum_j X_j/2\right)$, $P(Z = 2 \mid X) \propto \exp\left(-\alpha_{Z,2} - X_1 - X_2/2\right)$, and $P(Z = 3 \mid X) \propto \exp\left(-\alpha_{Z,3} - X_1/2 - X_2\right)$, where $M = 3$ and $\alpha_Z = (\alpha_{Z,1}, \alpha_{Z,2}, \alpha_{Z,3})$ is adaptively chosen to ensure the group proportions are approximately 0.5, 0.3, and 0.2, respectively; additional experiments with larger number of groups (e.g., $M = 10, 20$) are presented in Appendix B.

Algorithm 1 Surrogate-Assisted Group Clustered Conformal Inference (SAGCCI)

Input: Source data $\mathcal{D}_1 = \{(W_i, Z_i, Y_i, D_i = 1)\}_{i=1}^{n_{D1}}$, target data $\mathcal{D}_0 = \{(W_i, Z_i, Y_i = \text{NA}, D_i = 0)\}_{i=n_{D1}+1}^{N}$, and miscoverage level $\alpha \in (0, 1)$

◁ Preparation

Randomly split the data $\mathcal{D}_1 \cup \mathcal{D}_0$ into training and calibration folds $\mathcal{I}_1$ and $\mathcal{I}_2$.

◁ Training

Split $\mathcal{I}_1$ into sub-folds $\mathcal{I}_{11}$ and $\mathcal{I}_{12}$.
Fit score function $\widehat{R}(W, Y)$ on $\mathcal{I}_1 \cap \mathcal{D}_1$. ◁ e.g., quantile random forest
Fit propensity function $\widehat{e}_D(W)$ on $\mathcal{I}_1$. ◁ e.g., SuperLearner for binary outcomes
Obtain the initial threshold $\widehat{r}_\alpha^{\text{init}}$ on $\mathcal{I}_{11} \cap \mathcal{D}_1$ via the WCQR Algorithm from Lei & Candès (2021).
Fit conditional CDF $\widehat{m}(\widehat{r}_\alpha^{\text{init}}, W)$ and $\widehat{m}(\widehat{r}_\alpha^{\text{init}}, X)$ on $\mathcal{I}_{12} \cap \mathcal{D}_1$. ◁ e.g., SuperLearner for continuous outcomes

◁ Calibration

Compute $R_i = \widehat{R}(W_i, Y_i)$ for $i \in \mathcal{I}_2 \cap \mathcal{D}_1$.
Determine clustering number K by assessing R_i across Z_i. ◁ e.g., Elbow method or other herustics
Obtain the clustering mapping $\widehat{h}(Z)$ on $\mathcal{I}_2 \cap \mathcal{D}_1$. ◁ e.g., k-means
Compute the clustering indices $\mathcal{I}^{(k)} = \{i : \widehat{h}(Z_i) = k, i \in \mathcal{I}_2\}$ for $k = 1, \cdots, K$.
Compute the cluster-specific threshold $\widehat{r}_\alpha^k$ as the smallest values satisfying $\sum_{i \in \mathcal{I}_2 \cap \mathcal{I}^{(k)}} \psi_\alpha(V_i; \widehat{e}_D, \widehat{m}, \widehat{\widehat{m}}, r_\alpha^k) \geq 0$.

◁ Prediction

For a new data point $(W, Z, Y) \sim P_{D_0}$ from the target population, compute $C(W, \widehat{r}_\alpha^k)$, where $k = \widehat{h}(Z)$.

Output: prediction set $C(W, \widehat{r}_\alpha^k)$

The surrogate outcomes $S \in \mathbb{R}^2$ are modeled as $S \sim \mathcal{N}(\mathbf{1}_2, \sigma_S^2 I_2)$, where $\sigma_S = 1, 3, 5$. The variance σ_S^2 reflects the extent to which the surrogates S explain the variability in the primary outcome Y, after adjusting for the baseline covariates X. A larger σ_S^2 indicates higher predictiveness of the surrogates S for the primary outcomes Y.

Finally, the primary outcomes Y are generated as a multinomial variable with five levels. Specifically, for $y = 2, ..., 5$, the conditional probabilities are given by: $\frac{P(Y=y|X,S)}{P(Y=1|X,S)} = \exp\left(-\alpha_y - \sum_{j=1}^2 \frac{X_j}{2} - \sum_{j=1}^2 \frac{S_j}{2}\right)$, where $y = 1$ serves as the reference level, and α_y are adaptively chosen to ensure marginal probabilities of $P(Y = k) = (0.1, 0.2, 0.4, 0.15, 0.15)$ for $y = 1, \ldots, 5$.

The non-conformity scores for the categorical outcomes are derived using the nested prediction sets approach from Kuchibhotla & Berk (2023); more details are presented in the Appendix. Following Sesia & Candès (2020), we split 75% of the data into the first fold $\mathcal{I}_1$ for model training, and use the remaining 25% as the second fold $\mathcal{I}_2$ to construct prediction sets. We compare our proposed method (SAGCCI) with $K = 2$ clusters against four alternative approaches: 1) clustered conformal inference without leveraging surrogates (NoSURRO + CLUSTER); 2) surrogate-assisted group-wise conformal inference (SURRO + GROUP), which accounts for the covariate shifts to obtain the group-specific thresholds as in Gibbs et al. (2023); 3) surrogate-assisted standard conformal inference without accounting for protected groups (SURRO + STANDARD); 4) weighted conformal quantile regression (WCQR) from Lei & Candès (2021); and 5) conformalized fair quantile regression (CFQR) from Liu et al. (2022).

5.2. Evaluation metrics

Denote the second data fold $\mathcal{I}_2$ as the validation dataset: $\{(W_i, Z_i, Y_i)\}_{i \in \mathcal{I}_2}$. The empirical group-conditional coverage for group z is defined as: $\widehat{c}_z = \sum_{i \in \mathcal{I}^z \cap \mathcal{I}_2} \mathbf{1}(Y_i \in C(W_i))/|\mathcal{I}^z \cap \mathcal{I}_2|$, where $\mathcal{I}^z \cap \mathcal{I}_2$ represents the subset of validation data points belonging to group z, and $|\mathcal{I}^z \cap \mathcal{I}_2|$ is the number of such points.

To evaluate the quality of the prediction sets while incorporating considerations of group fairness, we use two primary metrics: 1) Average Size (AvgSize) of the prediction sets, defined as AvgSize $= \sum_{i \in \mathcal{I}_2} |C(W_i)|/|\mathcal{I}_2|$, where $|C(W_i)|$is the size of the prediction set for individual i; and 2) Average Group Coverage Gap (CovGap), defined as CovGap $= \sum_{z \in \mathcal{Z}} |\widehat{c}_z - (1 - \alpha)|/M$, where $M = |\mathcal{Z}|$ is the total number of groups.

The AvgSize metric assesses the precision of the prediction sets, where smaller values indicate more precise intervals. Meanwhile, the CovGap metric evaluates how closely the

group-conditional coverage aligns with the desired level $1-\alpha$. A smaller CovGap reflects better adherence to fairness goals across groups. Additionally, we evaluate the fraction of under-covered and over-covered groups in each experiment and provide the results in Appendix B; similar conclusions can be drawn from these results.

5.3. Results

First, we evaluate the benefits of leveraging surrogates to construct prediction sets. Figure 1 summarizes the performance metrics for the proposed clustered conformal inference methods, both with and without surrogates, across 500 Monte Carlo simulations for $\sigma_S = 1, 3, 5$, representing low, medium, and high surrogate effects, respectively. The WCQR method with surrogates produces more precise prediction sets compared to NoSurro + Cluster when surrogates are highly predictive of the primary outcomes (i.e., as σ_S increases). However, it fails to ensure valid group-conditional coverage, as indicated by its non-decreasing CovGap with increasing sample size. In contrast, SAGCCI consistently achieves the smallest average prediction set sizes while maintaining valid group-conditional coverage, evidenced by its empirically decreasing AvgSize and CovGap. Its efficiency gain over NoSurro + Cluster becomes more pronounced as σ_S increases. With larger σ_S, surrogates explain a greater proportion of the variability in the primary outcomes, increasing their predictiveness and improving the construction of prediction sets. The results align with the theoretical results presented in Corollary 4.5, where surrogates are leveraged in the most efficient way by the EIF to estimate the thresholds.

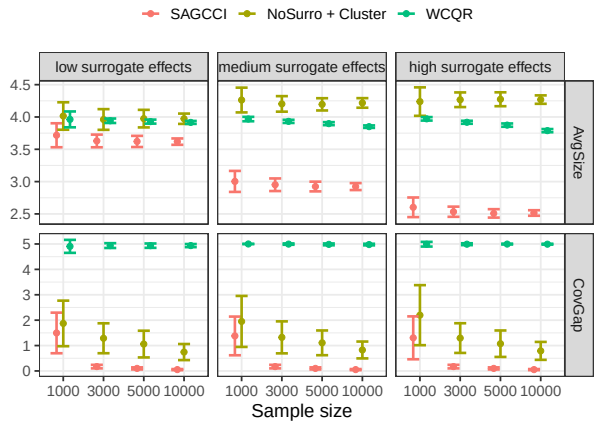

Figure 1. Comparison of AvgSize and CovGap for the considered methods. The error bar plot denote $\pm$ the standard errors.

Next, we investigate the impact of cluster mapping on empirical group-conditional coverages, with results presented in Figure 2. The Surro + Standard method achieves the smallest AvgSize across all settings by utilizing the entire dataset without accounting for protected groups. However, as expected, its CovGap does not decrease to zero, as it lacks guarantees for group-conditional coverage. In contrast, the Surro + Group method fairly ensures the desired group-conditional coverage but with a significantly larger AvgSize, indicating less precise prediction sets. The CFQR method focuses on learning a fair quantile function to construct the non-conformity score used to produce fair prediction sets. However, this approach may not generalize well to scores for categorical outcomes, as evidenced by its unsatisfactory performance under our problem setup. In contrast, our proposed SAGCCI method takes a balanced approach by leveraging the similarity in score distributions across protected groups to form larger clusters and estimate cluster-specific thresholds, which is more general and can be applied to any type of non-conformity scores. This strategy maintains high-quality and precise prediction sets comparable to Surro + Standard in terms of AvgSize, while effectively addressing disparities in group-conditional coverage for protected groups with a diminishing CovGap, similar to Surro + Group.

For example, when $\sigma_S = 3$, the proposed SAGCCI reduces AvgSize by over 60% on average compared to Surro + Group at $N = 1000$, while increasing the CovGap by only 7%. Compared to Surro + Standard, the AvgSizes are similar, but SAGCCI noticeably reduces the CovGap by over 80% at $N = 3000$. Even when surrogate outcomes are not available (i.e., $\sigma_S = 0$), cluster mapping remains beneficial, as seen by comparing the performance of NoSurro + Cluster against NoSurro + Standard and NoSurro + Group in Appendix B.

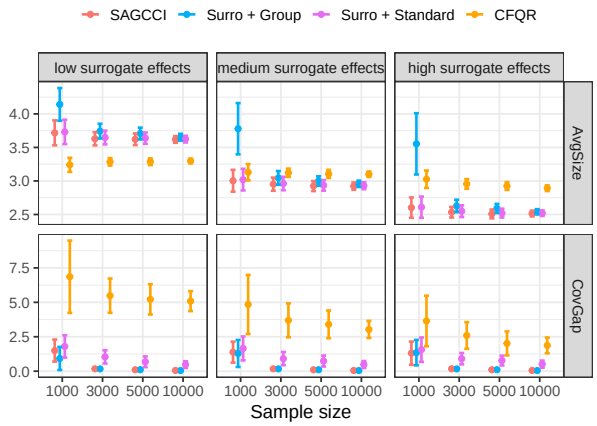

Figure 2. Comparison of AvgSize and CovGap for the considered methods. The error bar plot denote $\pm$ the standard errors.

6. Real-data application

We analyzed data from the Moderna COVE phase 3 COVID-19 vaccine efficacy trial, which randomized adults to receive two doses of mRNA-1273 or placebo at Days 1 and 29. Our per-protocol analysis included participants who received both injections without specified protocol violations. Participants were randomly sampled into the immunogenicity subcohort using stratified Bernoulli random sampling, with antibody markers measured at Days 1, 29, and 57. Our analysis focused on a cohort of 1418 individuals, including 46.0% underrepresented minorities ($n = 652$), defined as Blacks or African Americans, Hispanics or Latinos, American Indians or Alaska Natives, Native Hawaiians, and Pacific Islanders. The source population of non-minorities ($n = 766$) included all other races (e.g., White, Asian) and non-Hispanic ethnicity. Confounders included a standardized COVID-19 risk score built using machine learning and baseline covariates (e.g., sex, BMI, enrollment time), with fairness evaluations conducted across age-comorbidity stratification groups: (i) 65 or older, (ii) younger than 65 with comorbidities, and (iii) younger than 65 without comorbidities. The surrogate outcome was the Day 29 $\log_{10}$ nAb-ID50 titer, while the primary outcome was a discrete variable representing quintiles of the binding spike antibody level at Day 57.

We set the miscoverage level $\alpha = 0.05$ and employed a 75-25 train-test split, repeating the procedure 100 times to summarize results. Table 1 presents the results. Methods leveraging surrogate outcomes (SURRO + STANDARD, SURRO + GROUP, and SAGCCI) consistently outperformed those that did not, producing lower coverage gaps and shorter prediction interval sizes. Among these, our proposed SAGCCI achieves the best overall performance, with the shortest prediction intervals across all three age-comorbidity subgroups, and valid group-conditional coverage. This highlights the effectiveness of combining surrogate information with clustering to achieve both precision and fairness in individualized vaccine efficacy estimation.

Table 1. Comparison of methods: AvgSize, CovGap, empirical coverage, and width for age-comorbidity stratification groups

Method	AvgSize	CovGap	Coverage	Width
NOSURRO + STANDARD	4.46	4.39	(0.97, 0.87, 0.92)	(4.79, 4.20, 4.39)
SURRO + STANDARD	3.83	1.09	(**0.95**, 0.93, 0.96)	(3.82, 3.91, 3.76)
NOSURRO + GROUP	4.63	2.22	(0.97, 0.90, 0.95)	(4.81, 4.44, 4.63)
SURRO + GROUP	3.93	**0.14**	(**0.95, 0.95, 0.95**)	(3.80, 4.06, 3.94)
NOSURRO + CLUSTER	4.48	3.28	(0.94, 0.87, 0.94)	(4.62, 4.27, 4.53)
WCQR	4.05	5.24	(0.98,0.82,0.88)	(4.21, 3.82, 4.24)
SAGCCI (ours)	**3.61**	0.18	(**0.95, 0.95, 0.95**)	(**3.63, 3.55, 3.66**)

7. Discussion

We have introduced a novel surrogate-assisted group-clustered conformal inference (SAGCCI) framework for constructing fair and efficient prediction sets. SAGCCI substantially improves efficiency over existing group-wise conformal methods by producing tighter prediction sets while maintaining valid group-conditional coverage, making it well-suited for high-stakes applications such as clinical risk assessment and policy evaluation. A promising direction for future research is to extend the generalizability and transportability of these conformal inference methods across multiple data sources, such as multi-site randomized clinical trials or observational studies. This could involve relaxing the MAR assumption and addressing potential distributional shifts, including conditional shifts in outcomes and surrogates (Liu et al., 2024). Such extensions would be particularly valuable in settings where privacy constraints preclude the sharing of individual-level data, a common challenge in federated causal inference (Han et al., 2023; 2024; 2025; Guo et al., 2025; Xiong et al., 2023).

Software and Data

Our R code, complete with illustrative examples, is available at https://github.com/Gaochenyin/SurrConformalDR.

Acknowledgements

We thank the anonymous (meta)-reviewers of ICML 2025 for their helpful comments. We thank the COVE study participants and study team, other COVPN colleagues, and Moderna colleagues. This research was supported in part by the Administration for Strategic Preparedness and Response, Biomedical Advanced Research and Development Authority Contracts No. 75A50120C00034 (P3001 study) and No. 75A50122C00013 (Immune Assays) and the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health through grant UM1AI068635. The findings and conclusions herein are those of the authors and do not necessarily represent the views of the Department of Health and Human Services or its components. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Department of Health and Human Services.

Impact Statement

The proposed approach improves the efficiency of uncertainty quantification by leveraging surrogates, thereby reducing implementation costs in practice. Furthermore, it incorporates fairness to prevent models from discriminating against certain protected groups, aligning predictions with ethical and regulatory standards. Therefore, this approach builds trust and accountability in individualized decision-making across critical fields, ensuring compliance with societal and legal regulations.

References

Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M., and Burzykowski, T. An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. *Biometrics*, 72:669–677, 2016.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Athey, S., Chetty, R., and Imbens, G. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv Preprint arXiv:2006.09676*, 2020.

Chen, J. and Ritzwoller, D. M. Semiparametric estimation of long-term treatment effects. *Journal of Econometrics*, 237:105545, 2023.

Cheng, D., Ananthakrishnan, A. N., and Cai, T. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, 77:413–423, 2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Conlon, A., Taylor, J., Li, Y., Diaz-Ordaz, K., and Elliott, M. Links between causal effects and causal association for surrogacy evaluation in a gaussian setting. *Statistics in Medicine*, 36:4243–4265, 2017.

Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Elliott, M. R. Surrogate endpoints in clinical trials. *Annual Review of Statistics and its Application*, 10:75–96, 2023.

FDA. Accelerated approval of new drugs for serious or life-threatening illnesses. *Federal Register*, 57:58942–58960, 1992.

FDA. Accelerated approval program. https://www.fda.gov/drugs/nda-and-bla-approvals/accelerated-approval-program, 2021. Accessed: 2024-10-31.

Fleming, T. R. and DeMets, D. L. Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*, 125:605–613, 1996.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Gao, C., Gilbert, P. B., and Han, L. On the role of surrogates in conformal inference of individual causal effects. *arXiv preprint arXiv:2412.12365*, 2024.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Gilbert, P. B. and Hudgens, M. G. Evaluating candidate principal surrogate endpoints. *Biometrics*, 64:1146–1154, 2008.

Gilbert, P. B., Montefiori, D. C., McDermott, A. B., Fong, Y., Benkeser, D., Deng, W., Zhou, H., Houchens, C. R., Martins, K., Jayashankar, L., et al. Immune correlates analysis of the mrna-1273 covid-19 vaccine efficacy clinical trial. *Science*, 375:43–50, 2022.

Gilbert, P. B., Peng, J., Han, L., Lange, T., Lu, Y., Nie, L., Shih, M.-C., Waddy, S. P., Wiley, K., Yann, M., et al. A surrogate endpoint based provisional approval causal roadmap. *arXiv Preprint arXiv:2407.06350*, 2024.

Guan, L. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Guo, Z., Li, X., Han, L., and Cai, T. Robust inference for federated meta-learning. *Journal of the American Statistical Association*, pp. 1–16, 2025.

Han, L., Shen, Z., and Zubizarreta, J. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36: 70453–70482, 2023.

Han, L., Li, Y., Niknam, B., and Zubizarreta, J. R. Privacy-preserving, communication-efficient, and target-flexible hospital quality measurement. *The Annals of Applied Statistics*, 18(2):1337–1359, 2024.

Han, L., Hou, J., Cho, K., Duan, R., and Cai, T. Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, (just-accepted):1–25, 2025.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Imbens, G., Kallus, N., Mao, X., and Wang, Y. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkae095, 2024.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kallus, N. and Mao, X. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkae099, 2024.

Kallus, N., Mao, X., and Uehara, M. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25:1–59, 2024.

Krishnamoorthy, K. and Mathew, T. *Statistical tolerance regions: theory, applications, and computation*. John Wiley & Sons, 2009.

Kuchibhotla, A. K. and Berk, R. A. Nested conformal prediction sets for classification with applications to probation data. *The Annals of Applied Statistics*, 17:761–785, 2023.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76:71–96, 2014.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111, 2018.

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

Li, Y., Taylor, J. M., and Elliott, M. R. A bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*, 66:523–531, 2010.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Liu, M., Ding, L., Yu, D., Liu, W., Kong, L., and Jiang, B. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35:11561–11572, 2022.

Liu, Y., Levis, A., Normand, S.-L., and Han, L. Multi-source conformal inference under distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 31344–31382. PMLR, 2024.

Makhlouf, K., Zhioua, S., and Palamidessi, C. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642, 2021.

Prentice, R. L. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8: 431–440, 1989.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Sesia, M. and Candès, E. J. A comparison of some conformal quantile regression methods. *Stat*, 9:e261, 2020.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Van der Laan, M. J., Rose, S., et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Van der Vaart, A. W. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

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

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vovk, V., Nouretdinov, I., and Gammerman, A. On-line predictive linear regression. *The Annals of Statistics*, pp. 1566–1590, 2009.

Xia, E. and Wainwright, M. J. Prediction aided by surrogate training. *arXiv preprint arXiv:2412.09364*, 2024.

Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. Federated causal inference in heterogeneous observational data. *Statistics in Medicine*, 42 (24):4418–4439, 2023.

Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkae009, 2024.

Zhu, Y., Deng, Q., Huang, D., Jing, B., and Zhang, B. Clustering based on kolmogorov–smirnov statistic with application to bank card transaction data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70 (3):558–578, 2021.

A. Proofs

A.1. Proof of Theorem 4.4

In this section, we derive the EIF for the threshold r_α. Since the threshold can be group-(or cluster-)specific, the protected group Z is excluded from the observed data in the derivation. Let t denote the index for the parametric submodels $f_t(\mathcal{O})$ of the observed data (X, D, S, Y) with true density function evaluated at $t = t^*$, i.e., $f(\mathcal{O}) = f_{t^*}(\mathcal{O})$, where

$$\begin{aligned} f(\mathcal{O}) = f(X) \times e_D(X)^D \{1 - e_D(X)\}^{1-D} \\ \times f(S \mid D, X) \times f(Y \mid S, D, X), \end{aligned}$$

the observed score function is derived by the pathwise derivatives of $\log f_t(\mathcal{O})$ with respect to t, given by $s_t(\mathcal{O}) = s_t(Y \mid S, D, X) + s_t(S \mid D, X) + s_t(D \mid X) + s_t(X)$. The tangent space corresponding to this model is $\Lambda = \Lambda_X \oplus \Lambda_{D|X} \oplus \Lambda_{S|D,X} \oplus \Lambda_{Y|S,D,X}$, where Λ_X, $\Lambda_{D|X}$, $\Lambda_{S|D,X}$, and $\Lambda_{Y|S,D,X}$ are the mean square closures of the score vector of the submodels:

$$\begin{aligned} \Lambda_X &= \{\Gamma(X) : \int \Gamma(x) f(x) dx = 0\}, \\ \Lambda_{S|D,X} &= \{\Gamma(S, D, X) : \int \Gamma(s, D, X) f(S \mid D, X) ds = 0\}, \\ \Lambda_{Y|S,D,X} &= \{\Gamma(Y, S, D, X) : \int \Gamma(y, S, D, X) f(y \mid S, D, X) dy = 0\}, \end{aligned}$$

for any two arbitrary square-integrable measurable functions $a(X)$ and $b(X)$.

Let $u(R, r_\alpha) = \mathbf{1}(R \le r_\alpha \mid D = 0) - (1 - \alpha)$, and $r(t)$ be the corresponding $(1 - \alpha)$-quantile for the scores of the target population $R \mid D = 0$ under the parametric submodels, which satisfies $E_t[u\{R_1, r_1(t)\}] = 0$ with $r(t^*) = r_\alpha$. Next, we can show that

$$0 = \frac{\partial}{\partial t} E_t[u\{R, r(t)\}]\big|_{t=t^*} = \frac{\partial}{\partial t} E_t[u\{R, r(t^*)\}]\big|_{t=t^*} + \frac{\partial}{\partial r} E_{t^*}[u\{R, r(t^*)\}] \frac{\partial r(t)}{\partial t}\big|_{t=t^*},$$

which leads to

$$\frac{\partial r(t)}{\partial t}\big|_{t=t^*} = \left[-\frac{\partial}{\partial r} E_{t^*}\{u(R, r_\alpha)\}\right]^{-1} \frac{\partial}{\partial t} E_t\{u(R, r_\alpha)\}\big|_{t=t^*} \propto \frac{\partial}{\partial t} E_t\{u(R, r_\alpha)\}\big|_{t=t^*}.$$

Under Assumptions 4.1 and 4.2, we have

$$\begin{aligned} &\frac{\partial}{\partial t} E_t\{u(R, r_\alpha)\}\big|_{t=t^*} = \frac{\partial}{\partial t} P_t(R < r_\alpha \mid D = 0)\big|_{t=t^*} \\ &\propto \frac{\partial}{\partial t} E_{X,t}[E_{S,t}\{P_{Y,t}(R_1 < r_{\alpha,1} \mid D = 1, S, X) \mid X\} \mid D = 0]\big|_{t=t^*} \\ &\propto \frac{\partial}{\partial t} E_{X,t}[(1 - D) E_{S,t}\{P_{Y,t}(R < r_\alpha \mid D = 1, S, X) \mid X\}]\big|_{t=t^*} \\ &\propto \frac{\partial}{\partial t} E_{X,t}[P_t(D = 0 \mid X) E_{S,t}\{P_{Y,t}(R < r_\alpha \mid D = 1, S, X) \mid X\}]\big|_{t=t^*} \\ &= T_1 + T_2 + T_3 + T_4, \end{aligned}$$

which is constituted by five terms T_1, T_2, T_3, and T_4. We analyze these four terms separately,

$$\begin{aligned} T_1 &= \frac{\partial}{\partial t} E_{X,t}\{P(D = 0 \mid X) m(r, X)\}\big|_{t=t^*} = E_X[\{P(D = 0 \mid X) m(r, X) - (1 - D)(1 - \alpha)\} s(X)], \\ T_2 &= \frac{\partial}{\partial t} E_X\{P_t(D = 0 \mid X) m(r, X)\}\big|_{t=t^*} = -E[\{D - e_D(X)\} m_1(r, X) s(D \mid X)], \\ T_3 &= \frac{\partial}{\partial t} E_X[P(D = 0 \mid X) E_{S,t}\{\tilde{m}(r, W) \mid X\}]\big|_{t=t^*} = E\left[P(D = 0 \mid X)\left\{\tilde{m}(r, W) - m(r, X)\right\} s(S \mid D, X)\right], \end{aligned}$$

and

$$\begin{aligned}T_4 &= \frac{\partial}{\partial t} E_X[P(D=0 \mid X) E_S\{P_{Y,t}(R < r_\alpha \mid D=1, S, X) \mid X\}]\big|_{t=t^*} \\ &= \frac{\partial}{\partial t} E_X\left(P(D=0 \mid X) E\left[\frac{D\{\mathbf{1}(R \le r_\alpha) - \tilde{m}(r, W)\}}{e_D(X)} s(Y \mid S, D, X) \mid X\right]\right) \\ &= \frac{\partial}{\partial t} E\left[D\pi_D(X)\{\mathbf{1}(R \le r_\alpha) - \tilde{m}(r, W)\} s(Y \mid S, D, X)\right].\end{aligned}$$

Putting these terms together, we can show that

$$\frac{\partial r(t)}{\partial t}\big|_{t=t^*} \propto \frac{\partial}{\partial t} E_t\{u(R, r_\alpha)\}\big|_{t=t^*} = E\{\psi_\alpha(V; e_D, m, \tilde{m}, r_\alpha) s(Y, S, D, X)\},$$

and

$$\begin{aligned}\psi_\alpha(V; e_D, m, \tilde{m}, r_\alpha) &= (1-D)\{m(r_\alpha, X) - (1-\alpha)\} \\ &\quad + \{1 - e_D(X)\}\{\tilde{m}(r_\alpha, W) - m(r_\alpha, X)\} \\ &\quad + D\pi_D(X)\{\mathbf{1}(R < r_\alpha) - \tilde{m}(r_\alpha, W)\} \\ &= (1-D)\{m(r_\alpha, X) - (1-\alpha)\} + D\pi_D(X)\{\mathbf{1}(R \le r_\alpha) - m(r_\alpha, X)\} \\ &\quad + [D\pi_D(X) - \{1 - e_D(X)\}]\{m(r_\alpha, X) - \tilde{m}(r_\alpha, W)\},\end{aligned}$$

where $(1-D)\{m(r_\alpha, X) - (1-\alpha)\} \in \Lambda_X$, $\{1 - e_D(X)\}\{\tilde{m}(r_\alpha, W) - m(r_\alpha, X)\} \in \Lambda_{S|D,X}$, $D\pi_D(X)\{\mathbf{1}(R < r_\alpha) - \tilde{m}(r_\alpha, W)\} \in \Lambda_{Y|S,D,X}$. Thus, it completes the proof of Theorem 4.4 by the definition of efficient influence function.

A.2. Proof of Theorem 4.6

In this section, we establish asymptotic properties for the estimator of r_α, which is important for constructing prediction set with the user-defined coverage level $1-\alpha$. First, we establish the connection between the asymptotic coverage probability $P(R \le r_\alpha \mid D=0)$ and the EIF for r_α in Lemma A.1.

Under Lemma A.1, we can show that for the empirically estimated $\widehat{r}_\alpha$ using the second data fold $\mathcal{I}_2$, we have

$$\begin{aligned}&\mathbb{P}(R \le \widehat{r}_\alpha \mid D=0) - (1-\alpha) = \frac{\mathbb{P}_{\mathcal{I}_2}\{\widehat{\psi}(V; r_\alpha)\}}{P(D=0)} \\ &+ \frac{\mathbb{P}\{\widehat{\psi}(V; r_\alpha)\} - \mathbb{P}_{\mathcal{I}_2}\{\widehat{\psi}(V; r_\alpha)\}}{P(D=0)} + \frac{\mathbb{P}\{\widehat{\psi}(V; r_\alpha) - \psi(r_{\alpha,1}, W)\}}{P(D=0)} \ge 0 + I_1 + I_2.\end{aligned}$$

Here, we use the fact that $\mathbb{P}_{\mathcal{I}_2}\{\widehat{\psi}(V; r_\alpha)\} \ge 0$ by definition. The term I_1 is negligible if $\psi(V; r_\alpha)$ belongs to a Donsker class (Van der Vaart, 2000). Even if the Donsker condition is not met, the sample-splitting procedure in (Chernozhukov et al., 2018) can be used to assure that I_1 is negligible, where the first data fold $\mathcal{I}_1$ is used to estimate $\widehat{e}_D$, $\widehat{m}_a$ and $\widehat{\tilde{m}}$, and the second data fold $\mathcal{I}_2$ is used to compute r_α; see Lemma A.2 for more details on the bound of I_1.

The term I_2 is the second-order remainder term, which is bounded by the product of estimation error for the nuisance functions:

$$\begin{aligned}&\mathbb{P}\{\widehat{\psi}(V; r) - \psi(V; r)\} = \mathbb{P}\left[\{1 - e_D(X)\}\{\widehat{m}(r, X) - m(r, X)\}\right] \\ &+ \mathbb{P}\left[\{1 - \widehat{e}_D(X)\}\{\widehat{\tilde{m}}(r, W) - \widehat{m}(r, X)\}\right] + \mathbb{P}\left[e_D(X)\widehat{\pi}_D(X)\{\tilde{m}(r, W) - \widehat{\tilde{m}}(r, W)\}\right] \\ &= \mathbb{P}\left[\{1 - e_D(X)\}\{\widehat{m}(r, X) - m(r, X)\}\right] \\ &+ \mathbb{P}\left[\{1 - \widehat{e}_D(X)\}\{\widehat{\tilde{m}}(r, W) - \tilde{m}(r, W) + m(r, X) - \widehat{m}(r, X)\}\right] \\ &+ \mathbb{P}\left[e_D(X)\widehat{\pi}_D(X)\{\tilde{m}(r, W) - \widehat{\tilde{m}}(r, W)\}\right] \\ &= \mathbb{P}\left[\{\widehat{e}_D(X) - e_D(X)\}\{\widehat{m}(r, X) - m(r, X)\}\right] + \mathbb{P}\left[\{1 - \widehat{e}_D(X) - e_D(X)\widehat{\pi}_D(X)\}\{\widehat{\tilde{m}}_1(r, W) - \tilde{m}_1(r, W)\}\right] \\ &\lesssim \|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{m}(r, X) - m(r, X)\| + \|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{\tilde{m}}(r, W) - \tilde{m}(r, W)\|,\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality. Combining the bounds for I_1 and I_2 gives us

$$
\begin{aligned}
&P(R \le \widehat{r}_\alpha \mid D=0) \ge (1-\alpha) \\
&- C_0 \bar{e}_0 (\underline{e}_0^{-1} + m_0 + \underline{e}_0^{-1} \tilde{m}_0) \sqrt{\frac{\log(1/\delta)+1}{|\mathcal{I}_2|}} \\
&- C_1 \left\{ \|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{m}(r,X) - m(r,X)\| + \|\widehat{e}_D(X) - e_D(X)\| \cdot \sup_r \|\widehat{\tilde{m}}(r,W) - \tilde{m}(r,W)\| \right\},
\end{aligned}
$$

which completes the proof.

A.3. Proof of Corollary 4.5

By straightforward algebra, the EIF for r_α without surrogates is $\psi^r(V; r_\alpha) = (1-D)\{m(r_\alpha, X) - (1-\alpha)\} + D\pi_D(X)\{\mathbf{1}(R \le r_\alpha) - m(r_\alpha, X)\}$. The variance for estimating r_α without surrogates is $V^r_{\text{eff}} = \text{var}\{\psi^r(V; r_\alpha)\} = V_1^r + V_2^r + V_3^r + 2V_{12}^r - 2V_{13}^r - 2V_{23}^r$, where

$$
\begin{aligned}
&V_1^r = \text{var}\left[(1-D)\{m(r_\alpha, X) - (1-\alpha)\}\right], \\
&V_2^r = \text{var}\left[D\pi_D(X)\mathbf{1}(R \le r_\alpha)\right], \quad V_3^r = \text{var}\left\{D\pi_D(X) m(r_\alpha, X)\right\}, \\
&V_{12}^r = \text{cov}\left[(1-D)\{m(r_\alpha, X) - (1-\alpha)\}, D\pi_D(X)\mathbf{1}(R \le r_\alpha)\right] = 0, \\
&V_{13}^r = \text{cov}\left[(1-D)\{m(r_\alpha, X) - (1-\alpha)\}, D\pi_D(X) m(r_\alpha, X)\right] = 0, \\
&V_{23}^r = \text{cov}\left[D\pi_D(X)\mathbf{1}(R \le r_\alpha), D\pi_D(X) m(r_\alpha, X)\right] = V_3^r.
\end{aligned}
$$

Similarly, we can show that the variance for estimating r_α with surrogates is $V_{\text{eff}} = \text{var}\{\psi(V; r_\alpha)\} = V_1 + V_2 + V_3 + V_4 + 2V_{12} + 2V_{13} - 2V_{14} + 2V_{23} - 2V_{24} - 2V_{34}$, where

$$
\begin{aligned}
&V_1 = \text{var}\left[(1-D)\{m(r_\alpha, X) - (1-\alpha)\}\right], \\
&V_2 = \text{var}\left[\{1 - e_D(X)\}\{\tilde{m}(r_\alpha, W) - m(r_{\alpha,C}, X)\}\right] \\
&V_3 = \text{var}\left\{D\pi_D(X)\mathbf{1}(R \le r_\alpha)\right\}, \quad V_4 = \text{var}\left\{D\pi_D(X)\tilde{m}(r_\alpha, W)\right\}, \\
&V_{12} = V_{13} = V_{14} = 0, \\
&V_{23} = V_{24} \\
&\quad = \text{cov}\left[\{1 - e_D(X)\}\{\tilde{m}(r_\alpha, W) - m(X)\}, D\pi_D(X)\tilde{m}(r_\alpha, W)\right], \\
&V_{34} = \text{cov}\left[D\pi_D(X)\mathbf{1}(R \le r_\alpha), D\pi_D(X)\tilde{m}(r_\alpha, W)\right] = V_4.
\end{aligned}
$$

Therefore, we can verify that

$$
\begin{aligned}
&V^r_{\text{eff}} - V_{\text{eff}} = V_4 - V_3^r - V_2 \\
&= \text{var}\left\{D\pi_D(X)\tilde{m}(r_\alpha, W)\right\} - \text{var}\left\{D\pi_D(X) m_C(r_{\alpha,C}, X)\right\} - \text{var}\left[\{1 - e_D(X)\}\{\tilde{m}(r_\alpha, W) - m_C(r_{\alpha,C}, X)\}\right] \\
&= E\left\{D\pi_D^2(X)\tilde{m}^2(r_\alpha, W)\right\} - E\left\{D\pi_D^2(X) m_C^2(r_\alpha, X)\right\} - E\left[\{1 - e_D(X)\}^2\{\tilde{m}(r_\alpha, W) - m_C(r_\alpha, X)\}^2\right] \\
&= E\left[\frac{\{1 - e_D(X)\}^2}{e_D(X)}\{\tilde{m}^2(r_\alpha, W) - m_C^2(r_\alpha, X)\}\right] - E\left[\{1 - e_D(X)\}^2\{\tilde{m}^2(r_\alpha, W) - m_C^2(r_\alpha, X)\}\right] \\
&= E\left[\pi_D(X)\{1 - e_D(X)\}^2 \text{var}\{\tilde{m}_C(r_\alpha, X, S) \mid X\}\right],
\end{aligned}
$$

where $V_1^r = V_1$ and $V_2^r = V_3$. Hence, it completes the proof of Corollary 4.5.

A.4. Additional technical details

A.4.1. Proof of Lemma A.1

Lemma A.1. *Under Assumptions 4.1 and 4.2, the following holds for any EIF $\psi(V; r_\alpha)$:*

$$
P(R \le r_\alpha \mid D=0) = 1 - \alpha + \frac{E\{\psi(V; r_\alpha)\}}{P(D=0)}.
$$

On the one hand, we can show that

$$P(R \leq r_\alpha \mid D=0) = E\left\{\mathbf{1}(R \leq r_\alpha) \mid D=0\right\}. \tag{3}$$

On the other hand, we can prove that

$$\begin{aligned}
&E\{\psi(V; r_\alpha)\} = E[(1-D)\{m(r_\alpha, X) - (1-\alpha)\}] \\
&= E\left\{P(D=0 \mid X) P(R \leq r_\alpha \mid X)\right\} - P(D=0)(1-\alpha) \\
&= E\left\{(1-D) P(R \leq r_\alpha \mid X)\right\} - P(D=0)(1-\alpha) \\
&= P(D=0) E\left\{\mathbf{1}(R \leq r_\alpha) \mid D=0\right\} - P(D=0)(1-\alpha).
\end{aligned}$$

Combine it with (3), we have $P(R \leq r_\alpha \mid D=0) = 1-\alpha + E\{\psi(V; r_\alpha)\}/P(D=0)$.

A.4.2. Proof of Lemma A.2

Lemma A.2. *Under the regularity conditions (A1) and (A2), there exists some constant C_0 such that for any $\delta > 0$,*

$$P\left(|I_1| \leq \frac{C_0 \bar{e}_0 (\underline{e}_0^{-1} + m_0 + \underline{e}_0^{-1} \tilde{m}_0)}{P(D=1)} \sqrt{\frac{\log(1/\delta)+1}{N}} \mid \mathcal{I}_1\right) \geq 1-\delta.$$

The proof of Lemma A.2 is adapted from Theorem 3 in Yang et al. (2024). First, we expand the numerator of I_1 into the following four parts,

$$\begin{aligned}
&\mathbb{P}\{\widehat{\psi}(V; r_\alpha)\} - \mathbb{P}_{\mathcal{I}_2}\{\widehat{\psi}(V; r_\alpha)\} \\
&= \left[\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} D_i \widehat{\pi}_D(X_i) \mathbf{1}(R \leq r_\alpha) - \mathbb{P}\left\{D \widehat{\pi}_D(X) \mathbf{1}(R \leq r_\alpha)\right\}\right] && (4) \\
&+ \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left\{\widehat{e}_D(X_i) - D_i\right\} \widehat{m}_1(r_\alpha, X_i) - \mathbb{P}\left\{\widehat{e}_D(X) - D\right\} \widehat{m}_1(r_{\alpha,1}, X) && (5) \\
&+ \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \left[D_i \widehat{\pi}_D(X_i) - \{1 - \widehat{e}_D(X_i)\}\right] \widehat{\widetilde{m}}(r_\alpha, W_i) - \mathbb{P}\left[D \widehat{\pi}_D(X) - \{1 - \widehat{e}_D(X)\}\right] \widehat{\widetilde{m}}(r_\alpha, W) && (6) \\
&- \left\{\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} (1 - D_i) - P(D=0)\right\} (1-\alpha) && (7) \\
&= Rem_1(r_\alpha) + Rem_2(r_\alpha) + Rem_3(r_\alpha) + Rem_4.
\end{aligned}$$

Bound on $\sup_{r_\alpha} |Rem_1(r_\alpha)|$ We define a class of functions $\mathcal{F}_1$ by

$$\mathcal{F}_1 = \left\{f_r : f_r = D \pi_D(X) \mathbf{1}(R \leq r_\alpha), \forall r\right\}.$$

One can observe that $\forall f_r \in \mathcal{F}_1$, we have $|f_r| = |\widehat{\pi}_D(X)| \leq \pi_0$. Therefore, 1 is an envelope function of $\mathcal{F}_1$. Let $\|z\|_{\mathcal{F}}$ denote the supermum norm of z over a class of function $\mathcal{F}$, defined by $\|z\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |z(f)|$. For $\mathcal{O}_i = (X_i, A_i, S_i, Y_i, D_i)$ and any function f, we define

$$\mathbb{G}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [f(\mathcal{O}_i) - \mathbb{P}\{f(\mathcal{O})\}].$$

Applying Lemma A.3 with $s(a,d,w) = D$ and $h(w,y) = R$ produces $E\|\mathbb{G}_N\|_{\mathcal{F}_1} \leq C_1$ with C_1 being a universal constant. By McDiarmid's inequality, we can show that

$$P(\|\mathbb{G}_N\|_{\mathcal{F}_1} - E\|\mathbb{G}_N\|_{\mathcal{F}_1} \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^N c_i^2}\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^N 2^2/N}\right) = \exp\left(-\frac{t^2}{4\pi_0^2}\right),$$

where

$$c_i \leq \sup_{\mathcal{O}_i, \mathcal{O}_i'} \sup_r \sqrt{N} \left| \frac{1}{N} D_i \widehat{\pi}_D(X_i) \mathbf{1}(R_{1,i} < r) - \frac{1}{N} D_{i'} \widehat{\pi}_D(X_i') \mathbf{1}(R_{1,i'} < r) \right| \leq \frac{2\pi_0}{\sqrt{N}}.$$

Substituting the expectation bound to the inequality gives us

$$P\left(\|\mathbb{G}_N\|_{\mathcal{F}_1} \geq C_2 \pi_0 \sqrt{1 + \log\left(\frac{1}{\delta}\right)} \right) \leq \delta, \tag{8}$$

for some constant C_2.

Bound on $\sup_{r_\alpha} |Rem_2(r_\alpha)|$ Similar to the bound on $\sup_{r_\alpha} |Rem_1(r_\alpha)|$, a class of function $\mathcal{F}_2$ is defined by $\mathcal{F}_2 = \{f_r : f_r = \{\widehat{e}_D(X) - D\}\, \widehat{m}(r_\alpha, X), \forall r\}$, where

$$\begin{aligned} f_r &= \{\widehat{e}_D(X) - D\}\, \widehat{m}(r_\alpha, X) \\ &= \{\widehat{e}_D(X) - D\} \int_0^{m_0} \mathbf{1}\{\widehat{m}(r_\alpha, X) \geq u\} du \\ &= \int_0^{m_0} \{\widehat{e}_D(X) - D\}\, \mathbf{1}\{r_\alpha \geq h(X, u)\} du, \end{aligned}$$

where the first equality holds by $\widehat{m}(r_\alpha, X) = \int_0^{m_0} \mathbf{1}\{\widehat{m}(r_\alpha, X) \geq u\} du$, and the second equality holds as $\widehat{m}(r_\alpha, X)$ is monotonously increase in r_α. By Lemma A.3 and McDiarmid's inequality, we have

$$E\|\mathbb{G}_N\|_{\mathcal{F}_2} \lesssim \int_0^{m_0} du \lesssim m_0, \quad P(\|\mathbb{G}_N\|_{\mathcal{F}_2} - E\|\mathbb{G}_N\|_{\mathcal{F}_2} \geq t) \leq \exp\left(-\frac{t^2}{4m_0^2}\right).$$

Substituting the expectation bound back gives us

$$P\left(\|\mathbb{G}_N\|_{\mathcal{F}_2} \geq C_3 m_0 \sqrt{1 + \log\left(\frac{1}{\delta}\right)} \right) \leq \delta, \tag{9}$$

for some constant C_3.

Bound on $\sup_{r_\alpha} |Rem_3(r_\alpha)|$ Similarly, we first define the class of function $\mathcal{F}_3$ as

$$\mathcal{F}_3 = \left\{ f_r : [D\widehat{\pi}_D(X) - \{1 - \widehat{e}_D(X)\}]\, \widehat{\widehat{m}}(r_\alpha, W) \right\}.$$

The supermum norm of $\mathbb{G}_N$ over $\mathcal{F}_3$ and the bound of $\|\mathbb{G}_N\|_{\mathcal{F}_3}$ can be obtained by Lemma A.3 and McDiarmid's inequality, respectively:

$$E\|\mathbb{G}_N\|_{\mathcal{F}_3} \lesssim \int_0^{\tilde{m}_0} \pi_0 du \lesssim \pi_0 \tilde{m}_0, \quad P(\|\mathbb{G}_N\|_{\mathcal{F}_3} - E\|\mathbb{G}_N\|_{\mathcal{F}_3} \geq t) \leq \exp\left(-\frac{t^2}{4\pi_0^2 \tilde{m}_0^2}\right),$$

which yields

$$P\left(\|\mathbb{G}_N\|_{\mathcal{F}_3} \geq C_4 \pi_0 \tilde{m}_0 \sqrt{1 + \log\left(\frac{1}{\delta}\right)} \right) \leq \delta, \tag{10}$$

for some constant C_4.

Bound on $|Rem_4|$ The random variables $(1 - A_i)$ is i.i.d., and Hoeffding inequality gives us

$$P\left(\left\{ \frac{1}{N} \sum_{i=1}^{N} (1 - D_i) - P(D = 0) \right\} \geq t \right) \leq \exp\left(-\frac{2t^2}{N}\right)$$

which leads to

$$P\left(Rem_4 \geq (1-\alpha)\sqrt{\frac{1}{2N}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta. \tag{11}$$

Combining (8), (9), (10), and (11) together with the help of union bound, we can show that with probability larger than $1-\delta$,

$$\sup_{r_\alpha} |Rem_1(r_\alpha) + Rem_2(r_\alpha) + Rem_3(r_\alpha) + Rem_4| \lesssim (\pi_0 + m_0 + \pi_0\tilde{m}_0)\sqrt{\frac{\log(1/\delta)+1}{|\mathcal{I}_2|}},$$

which completes the proof for the bound of I_1.

A.4.3. Additional lemmas and theorems

Lemma A.3 (Lemma 8, Yang et al. (2024))**.** *There exists a universal constant C such that for any functions $s(a,d,w) \in [-\kappa_0, \kappa_0]$ and $h(w,y)$,*

$$E\left[\sup_r |\mathbb{G}_N[s(a,d,w)\mathbf{1}\{h(w,y) \leq r\}]|\right] \leq C\kappa_0.$$

A.5. Proof of Lemma 4.3

Denote $\mathcal{Z}^{(k)} = \{z \in \mathcal{Z} : \widehat{h}(z) = k\}$ be the set of protected groups that the clustering algorithm $\widehat{h}$ assign to cluster k. Let $R^{(k)}$ be the non-conformity scores for cluster k and R^z be the non-conformity scores for group z. Since we assume that the KS distance between the score distribution for each pair of the cluster k is bounded by ϵ, it follows that $\text{KS}(R^z, R^{(k)}) \leq \epsilon$ by the triangle inequality as $R^{(k)}$ follows a mixture distribution of the group-specific R^z such that $z \in Z^{(k)}$.

By the definition of KS distance, it implies $|P(R(W,y) \leq r_\alpha^k \mid Z = z) - P(R(W,y) \leq r_\alpha^k \mid Z \in \mathcal{Z}^{(k)})| \leq \epsilon$, where r_α^k is the cluster-specific threshold for cluster k. Since performing conformal inference separately for each cluster guarantees the coverage, we have

$$|P(R(W,y) \leq r_\alpha^k \mid Z = z) - P(y \in C(W; r_\alpha^k) \mid Z \in \mathcal{Z}^{(k)})| \leq \epsilon,$$

where $P(y \in C(W; r_\alpha^k) \mid Z \in \mathcal{Z}^{(k)}) \geq 1-\alpha$. Therefore, we have $P(R(W,y) \leq r_\alpha^k \mid Z = z) \geq 1-\alpha-\epsilon$, which completes the proof by integrating (W, Z, Y) over the target population.

B. Additional simulation details

Guidance on selecting the number of clusters The number of clusters K should be pre-specified, and we provide a simple heuristic for selecting this parameter in practice. On the one hand, each cluster should have a sufficiently large sample size to ensure efficient estimation of the thresholds. On the other hand, K should not be too large; otherwise this will reduce the method to group-wise conformal inference.

In particular, we recommend targeting an average of 100 subjects per cluster in our study, while Ding et al. (2024) advocated for 150 subjects per cluster in their practical applications. Therefore, K is chosen as the smallest number of clusters such that each cluster has at least 100 subjects, which is adopted throughout our simulations and real-data analyses. Other human-in-the-loop strategies can be adopted to avoid overly clustering on non-clusterable groups, such as the Elbow method or other information criterion (e.g., AIC, BIC) in a cross-validation setting.

Details of nonconformity scores for categorical primary outcomes Let $F(W;t) = \{y : P_Y(y \mid W) \leq t\}$ denote a nested sequence of sets constructed using the first data fold, $\mathcal{I}_1$. The corresponding non-conformity scores $R(W,Y)$ for the source data are defined as $R(W,Y) = \inf\{t : y \in F(W;t)\}$, where the conditional probabilities $P_Y(y \mid W)$ are estimated using multinomial logistic regression from the NNET package. To construct the scores for the target data, where the primary outcomes are unobserved, we fit a model on the scores $R(W_i, Y_i)$ from the source data $i = 1, \cdots, n_{D1}$ on W_i, and predict the scores for the target data. Finally, the semiparametric efficient estimator $\widehat{r}_\alpha^k$ for cluster k is computed as the smallest value satisfying the condition $\sum_{i \in \mathcal{I}_2 \cap \mathcal{I}^{(k)}} \psi_\alpha(V_i; \widehat{e}_D, \widehat{m}, \widehat{\widetilde{m}}, \widehat{r}_\alpha^k) \geq 0$, where the nuisance models $\widehat{m}$, $\widehat{\widetilde{m}}$, and $\widehat{e}_D$ are estimated from $\mathcal{I}_1$. The models are all fitted using SUPERLEARNER (Van der Laan et al., 2007), with Random Forests and generalized linear models as base learners.

Fraction of under-coverage and over-coverage The fraction of protected groups that are under-covered and over-covered are evaluated as well. We define the fraction of under-covered (FracUnderCov) and over-covered (FracOverCov) groups by: FracOverCov $= \sum_{z\in\mathcal{Z}} \mathbf{1}(\widehat{c}_z > 1-\alpha+0.01)/M$ and FracUnderCov $= \sum_{z\in\mathcal{Z}} \mathbf{1}(\widehat{c}_z < 1-\alpha-0.01)/M$, where the value 0.01 is included to account for finite-sample randomness and $\widehat{c}_z$ is the empirical group-conditional coverage for group z.

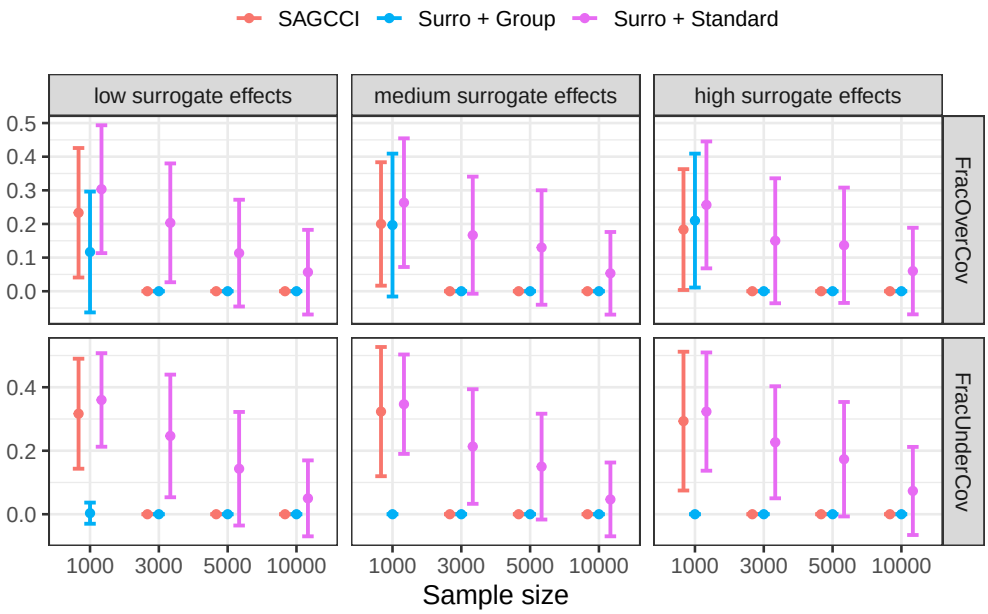

Figure 3. Comparison of FracOverCov and FracUnderCov for the considered surrogate-assisted methods. The error bar plot denotes $\pm$ the standard errors.

Figure 3 illustrates the results for the considered surrogate-assisted methods. It can be observed that both SURRO + GROUP and SAGCCI exhibit well-controlled fractions of under-coverage and over-coverage as the sample size increases. However, SURRO + GROUP achieves the low FracOverCov and FracUnderCov by producing larger prediction sets on average, which can be observed in Figure 2. Moreover, SURRO + STANDARD continues to exhibit non-negligible fractions of under-coverage or over-coverage, even with large sample sizes, as it is unable to guarantee group-conditional coverage.

Larger number of protected groups and clusters To evaluate the impact of a larger number of protected groups, we generate the protected group variable Z by randomly sampling from $\{1, \cdots, M\}$, where the number of groups M is set to 10 or 20. The baseline covariates X and the surrogate outcomes S are generated in the same way as described in the main simulation studies. Next, the primary outcomes Y are generated as a multinomial variable with five levels, where the conditional probabilities are given by: $\frac{P(Y=y|X,Z,S)}{P(Y=1|X,Z,S)} = \exp\left(-\alpha_y - \sum_{j=1}^{2} \frac{Z}{M}\frac{X_j}{2} - \sum_{j=1}^{2}\frac{S_j}{2}\right)$, with the same marginal probabilities $P(Y=k) = (0.1, 0.2, 0.4, 0.15, 0.15)$ as in the main paper.

Figure 4 presents the results for a fixed number of clusters, $K = 5$. As the number of groups increases, the performance of SURRO+GROUP noticeably degrades in both metrics (AvgSize and CovGap), even when $N \geq 5000$ for $M = 20$. This is because each group contains fewer samples, making it harder to estimate non-conformity scores accurately. In contrast, SAGCCI adopts a more balanced approach, achieving an AvgSize comparable to SURRO+STANDARD while also maintaining a smaller CovGap as the sample size increases. Furthermore, in the absence of surrogate outcomes, NOSURRO+CLUSTER, the non-surrogate-assisted version of SAGCCI, still delivers strong performance by effectively balancing AvgSize and CovGap.

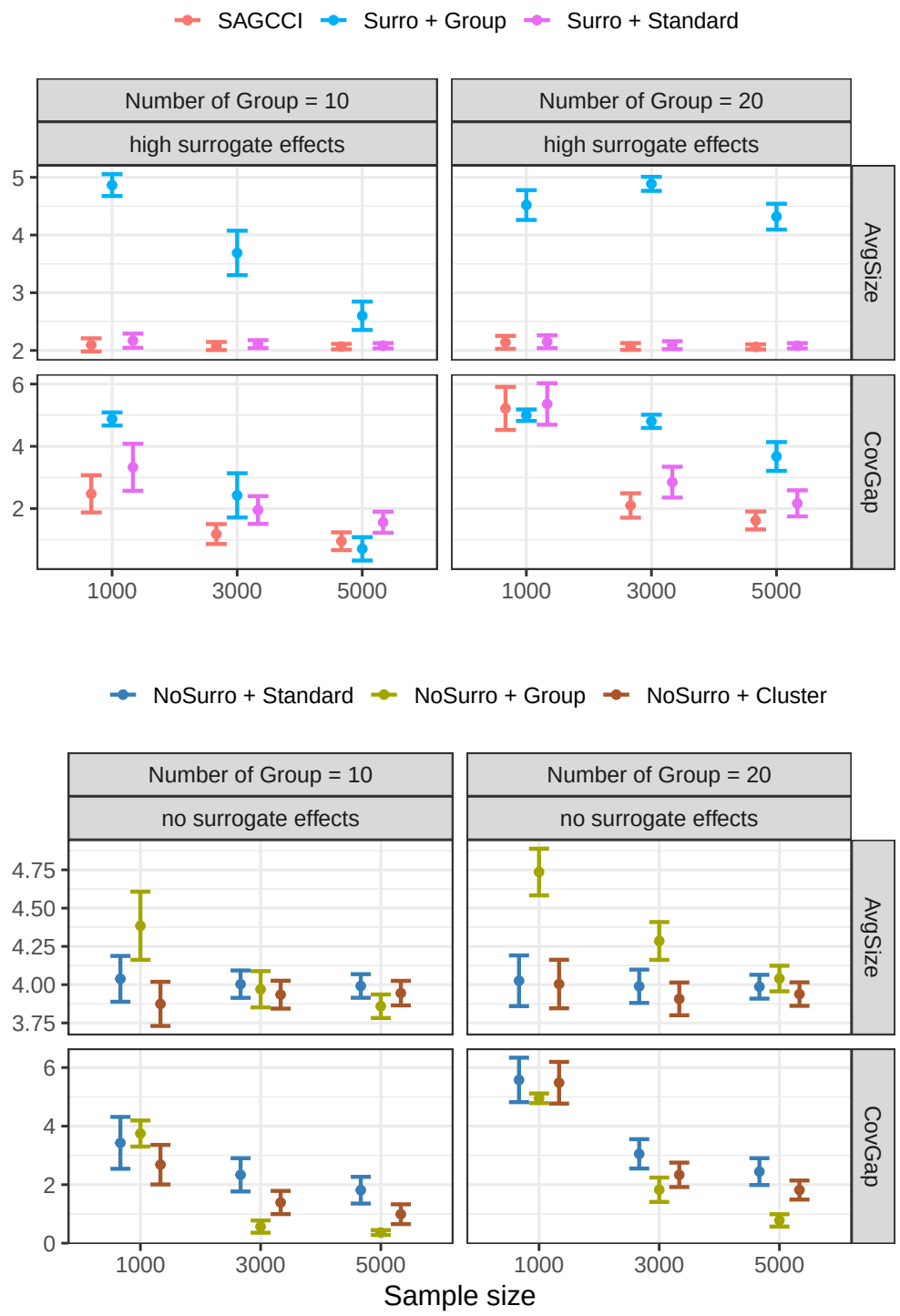

Figure 4. Comparison of AvgSize and CovGap for the considered methods for a larger number of groups and clusters. The error bar plot denotes $\pm$ the standard errors.