Stochastic Grouping Conformal Prediction for Effective Subgroup Reliability

Meihui Zhong[1]*, **Wenxin Tai**[1], **Ting Zhong**[1], **Fan Zhou**[1,2]†

[1] University of Electronic Science and Technology of China
[2] Intelligent Digital Media Technology, Key Laboratory of Sichuan Province

Abstract

Conformal prediction offers a distribution-free coverage guarantee, making it especially attractive for clinical applications. Standard conformal prediction, however, provides such guarantees only at the population level, and its prediction sets can exhibit coverage disparities across clinically important subgroups. A natural remedy is to calibrate within predefined groups. However, this can require access to sensitive subgroup attributes and is prone to a worst-group bottleneck: protecting the most difficult subgroup can inflate prediction sets for all, increasing cognitive burden on decision makers. To this end, we propose Stochastic Grouping Conformal Prediction (SGCP), a conformal framework for subgroup-reliable uncertainty quantification. It learns a stochastic grouping map that allows each sample to draw calibration information from others with similar calibration behavior, yielding a local score law that boosts reliability across subpopulations. We prove that SGCP retains the standard coverage guarantee. Experiments on synthetic and real-world benchmarks show that it consistently reduces subgroup coverage gaps while achieving smaller or comparable prediction set sizes relative to existing baselines. Our code is publicly available at `https://github.com/MHZhong1225/Restricted-calibration-CP`.

1 Introduction

Conformal prediction (CP) has emerged as an attractive framework for medical AI, because it converts any predictive model into a prediction set with finite-sample coverage guarantees under minimal assumptions [19, 5, 18]. Consider training data $\{(X_i, Y_i)\}_{i=1}^n$ and a future test point (X_{n+1}, Y_{n+1}), where each feature $X_i \in \mathcal{X}$ lies in the feature space and response $Y_i \in \mathcal{Y}$ lies in the label space. The goal is to construct a prediction set $\widehat{C}_n(X_{n+1}) \subseteq \mathcal{Y}$ that contains the unknown test response Y_{n+1} with high probability. The appeal of distribution-free prediction is that this guarantee can be obtained without imposing strong assumptions on the underlying data distribution [27, 24, 1], namely,

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\} \geq 1 - \alpha. \tag{1}$$

While this guarantee is highly appealing, it has an important limitation. The coverage in Eq. (1) is averaged over the calibration data and the future test point, thus providing validity only at the population level [10, 23]. In clinical settings, this means that a nominal 90% coverage guarantee may hold on average over all patients, while still failing to provide equally reliable uncertainty estimates for particular demographic groups, disease subtypes, or other clinically meaningful subpopulations.

This concern is tied to fairness in healthcare, where the costs of unreliable uncertainty often fall disproportionately on underrepresented patient populations [9, 4]. Differences in prevalence, data

*`mh.zhong1225@outlook.com`
†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

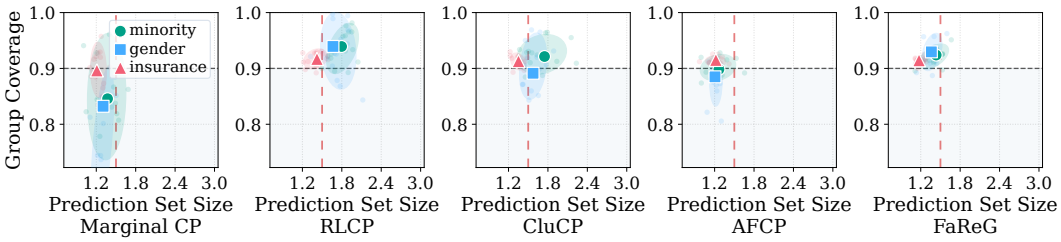


Figure 1: Tradeoff between subgroup coverage and prediction set size on MIMIC-IV [13] ($1 - \alpha = 0.9$). Results are shown for RLCP [12], CluCP [8], AFCP [32], and FaReG [30]. The red dashed line marks half of the total number of classes. Several fairness-aware CP methods improve sensitive-group coverage, but often do so with much larger prediction sets, making them less informative in practice.

quality, and feature distributions across populations can induce distinct error patterns and nonconformity score distributions [16]. Thus, population-level calibration is prone to masking subgroup failures by averaging under-coverage in some groups with over-coverage in others. In medical applications, this imbalance matters, as what we care about is not only whether uncertainty is valid on average, but whether it remains reliable for the subgroups for whom prediction errors are most costly.

We empirically observe exactly this tension, as shown in Figure 1. Standard conformal prediction achieves the desired overall coverage, but still exhibits lower coverage on several clinically important subpopulations, suggesting that marginal validity may 'average out' group-specific uncertainty. A natural response is to calibrate separately within each group. However, methods [22, 30] that target finer grained group reliability often become increasingly conservative, resulting in larger prediction sets. Such enlarged sets may be less informative in practice and may even hinder downstream decision-making by imposing excessive cognitive burden [31, 7, 6]. One may go even further and calibrate at the individual-sample level, as in localized conformal prediction [12, 22]. Yet our results show that localization, while sometimes improving coverage on subgroups, often incurs substantially larger prediction sets and, more importantly, greater instability across random seeds and data splits.

In sum, these observations pose two coupled challenges. First, the heterogeneity that matters for calibration is often latent: it may *not be* directly observable at inference, nor align with pre-specified sensitive groups. Second, the groups that most need tailored calibration are often those with *limited* sample sizes, making fine grained localized calibration conservative and statistically unstable [8]. More broadly, we seek a practical middle ground between global calibration, which can wash out subgroup-specific uncertainty, and calibration strategies that rely too heavily on known groups or local neighborhoods, which can become conservative or unstable when calibration data are limited.

Motivated by this gap, we propose **stochastic grouping conformal prediction** (SGCP), a fairness-aware conformal prediction method that improves subgroup reliability without relying on known sensitive attributes. Rather than asking coverage to hold for each individual test point, we relax the requirement to a weaker but still meaningful sense – coverage over samples that exhibit similar calibration behavior. In other words, exact 90% coverage may be unattainable for one patient in isolation, yet it may still be possible to achieve more reliable coverage on patients whose scores reflect similar uncertainty patterns and calibration difficulty.

To achieve this, SGCP learns a latent calibration structure that organizes samples according to uncertainty-relevant score behavior. Atop this structure, it learns a *stochastic grouping map* that determines how each sample should draw calibration information from a small set of latent calibration components. This yields a sample-adaptive local score law: calibration is tailored to each sample, but estimated through structured sharing rather than from a small group or neighborhood alone. Statistically, this design strikes a balance between flexibility and stability. It is flexible enough to capture heterogeneous score behavior across samples, yet structured enough to be reliably estimated from finite data while preserving conformal validity. As a result, SGCP can improve subgroup reliability beyond global calibration without relying on sparse subgroup calibration data. Empirically, across fairness-sensitive clinical and public prediction tasks, SGCP improves coverage on clinically meaningful subgroups while maintaining valid overall coverage and efficient prediction sets.

Our contributions are summarized as follows. First, we propose SGCP, a novel fairness-aware conformal prediction method. It learns a local score law through stochastic grouping, thereby approximating the score distribution of test points. This improves subgroup reliability without relying on rigid predefined groups. Second, we prove that SGCP retains finite-sample marginal validity. We further show that subgroup coverage gaps decrease when the learned local score law better captures the score behavior of meaningful subpopulations. Third, through experiments on synthetic and medical datasets, we demonstrate that SGCP achieves a favorable tradeoff among overall coverage, subgroup reliability, and prediction set efficiency. We expect SGCP to enable more effective and trustworthy decision-making in high-risk medical applications. Code and supplementary materials are included in the anonymous supplementary upload.

2 Preliminaries

We consider multiclass conformal classification with an exchangeable dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y} = \{1, \dots, C\}$. Let $f_\theta : \mathcal{X} \to \Delta^{C-1}$ be a base classifier, with $f_\theta^y(x)$ denoting the predicted probability of class y. We use the nonconformity score $s(x, y) = 1 - f_\theta^y(x)$. For an observed example (X_i, Y_i), we write $s_i = s(x_i, y_i)$, and it measures how inconsistent a point is with the true label. Given a training set $\mathcal{D}_{\text{tra}}$ and a calibration set $\mathcal{D}_{\text{cal}}$, we are interested in constructing, for a test point X_{n+1}, a confidence set $\widehat{C}(x) \subseteq \mathcal{Y}$ such that $\mathbb{P}\{Y_{n+1} \in \widehat{C}(X_{n+1})\} \geq 1 - \alpha$, for a miscoverage level α, without distributional assumptions beyond exchangeability.

In fairness-sensitive applications, the central issue is not merely whether the prediction set is valid on average, but rather how coverage and set size are distributed across heterogeneous subpopulations. Ideally, one desire distribution-free *conditional* coverage – that is,

$$\mathbb{P}\{Y \in \widehat{C}(x) \mid X = x\} \geq 1 - \alpha \qquad P_X\text{-a.e. } x. \tag{2}$$

which would mitigate uneven coverage across patient demographics, disease patterns, and other meaningful groups. However, non-trivial distribution-free conditional coverage is impossible, since finite samples cannot ensure tight, valid prediction sets for arbitrary conditioning events [10, 12].

We typically classify existing approaches into two broad types. **(i) Group-wise calibration.** A natural response to subgroup miscoverage is to calibrate separately across groups [19, 28], a type we term *Group-wise calibration* CP. If $G(x) \in \{1, \dots, K\}$ denotes a group label, this yields a group-specific threshold $\hat{q}_k = \text{Quantile}\Big(\{s_i : (x_i, y_i) \in \mathcal{D}_{\text{cal}},\ G(x_i) = k\},\ 1 - \alpha\Big)$, and outputs

$$\widehat{C}_{\text{GW}}(x) = \{y \in \mathcal{Y} :\ s(x, y) \leq \hat{q}_{G(x)}\}. \tag{3}$$

In principle, Eq. (3) reduces the mismatch induced by a single global threshold. In practice, however, one issue is that the relevant groups are often not known a priori, and may not align with pre-specified sensitive attributes. Another is that finer grouping fragments the calibration set, leaving some groups with too few samples for calibration, and making the empirical quantiles $\hat{q}_k$ statistically unstable [8].

This group-wise perspective also includes methods that learn or adapt groups before calibration. AFCP [32], for example, considers groups defined by sensitive attributes and makes the group selection itself adaptive. Let $\phi(x, \mathcal{A})$ denote the projection of x onto a selected sensitive-attribute set $\mathcal{A}$. For instance, a subgroup "males with a bachelor's degree" can be written as $\phi(x, \{k, \ell\}) \in [M_k] \times [M_\ell]$. For each test point, AFCP selects either one sensitive attribute or none, $\hat{A}(X_{n+1}) \in \{\emptyset, \{1\}, \dots, \{K\}\}$, and targets $\mathbb{P}\Big[Y_{n+1} \in \hat{C}(X_{n+1}) \mid \phi\big(X_{n+1}, \hat{A}(X_{n+1})\big)\Big] \geq 1 - \alpha$. This reduces to marginal coverage when $\hat{A}(X_{n+1}) = \emptyset$. Operationally, AFCP combines a marginal conformal set with an equalized-coverage set for each candidate label under the selected attribute. Although less conservative than exhaustive calibration over all attribute combinations, AFCP remains tied to discrete sensitive attributes. Its repeated leave-one-out calibration on pseudo-test points can also be computationally expensive and noisy in sparse-group regimes.

FAREG [30] further improves group discovery by learning groups in a latent representation space rather than enumerating them from raw input features. This expands the class of detectable subgroups, including those defined by nonlinear feature combinations. However, FAREG still frames fairness as subgroup discovery: it identifies explicit subgroups with coverage disparity and enforces coverage over them. It is therefore most suitable when heterogeneity is well captured by identifiable subgroups. It may, however, become conservative in some cases.

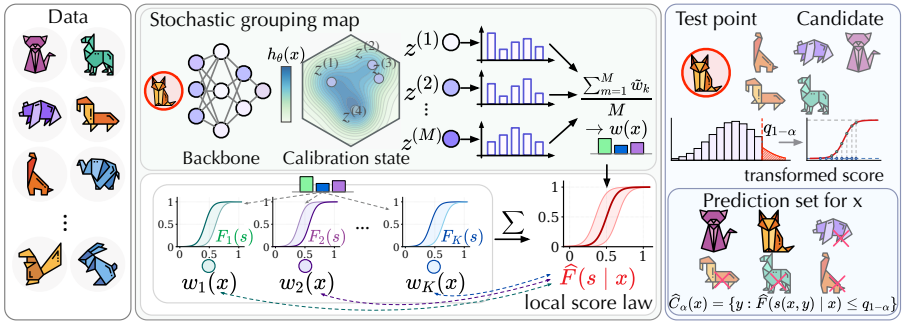


Figure 2: Overview of SGCP. SGCP learns a local score law by mapping x to a stochastic grouping map $w(x)$ over latent calibration components $\{F_k\}_{k=1}^K$. Candidate labels are transformed by this local law and kept when their calibrated scores fall below the conformal threshold.

(ii) Localized calibration. Localized calibration, as in RLCP [12], avoids explicit group partitions by using nearby samples as the reference set for each test point. RLCP argues that heterogeneity relevant to calibration may be highly sample-specific. Let $\mathfrak{X}_x = \{\mathcal{X} \in \mathfrak{X} : x \in \mathcal{X}\}$ denote the candidate neighborhoods containing x. A localized threshold can be written as $\hat{q}_{\mathrm{loc}}(x) = \sup_{\mathcal{X}\in\mathfrak{X}_x} \mathrm{Quantile}\Big(\{s_i : (x_i, y_i) \in \mathcal{D}_{\mathrm{cal}},\ x_i \in \mathcal{X}\},\ 1-\alpha\Big)$, with prediction set

$$\widehat{C}_{\mathrm{loc}}(x) = \{y \in \mathcal{Y} :\ s(x,y) \le \hat{q}_{\mathrm{loc}}(x)\}. \tag{4}$$

Compared with group-wise calibration, Eq. (4) avoids rigid group boundaries and adapts to the region around the test point. Yet the supremum over candidate neighborhoods requires the prediction set to cover the worst local subset containing x, which may be overly conservative and yield unnecessarily large sets, particularly when calibration samples are scarce. Moreover, locality is typically defined via feature space similarity, which need not align with the actual error structure most relevant for calibration. RLCP may introduces additional randomness into the prediction intervals, leading to inconsistent outputs across independent runs. Other related works are discussed in Appendix F.

The preceding discussion suggests that the key challenge is to improve subgroup reliability without losing statistical stability or producing overly large prediction sets. We therefore seek a method that can capture calibration heterogeneity while promoting informative prediction sets.

3 Method

3.1 From subgroup reliability to local score laws

Our starting point is that conformal calibration is ultimately governed by the distribution of the nonconformity score. Accordingly, we take the sample-specific score distribution $F_x^\star(s) := \mathbb{P}(S \le s \mid X = x)$, where S denotes the nonconformity score, as the oracle object for reliable calibration to each subgroup. If we could estimate $F_x^\star$ accurately, then each sample could be calibrated according to its own local score law. That said, the difficulty is that directly learning an unconstrained conditional distribution $F_x^\star$ for every x is data-hungry and unstable in finite samples, especially when the relevant heterogeneity is latent and high-dimensional. A single test point rarely provides enough local calibration evidence for reliable estimation. This motivates the central question behind SGCP: can other samples help a target point estimate its local score distribution without resorting to a single global calibration? SGCP addresses this by replacing the unconstrained family $\{F_x^\star : x \in \mathcal{X}\}$ with a structured mixture family. Concretely, SGCP approximates the oracle local score law by

$$F_x^\star(s) \approx \widehat{F}(s \mid x) = \sum_{k=1}^{K} w_k(x)\, F_k(s), \tag{5}$$

where $F_1, \ldots, F_K$ are *latent calibration components* and $w(x)$ is a sample-related *stochastic grouping map*. Thus, each sample has its own effective local score law, but that law is not estimated from scratch. It is represented as a weighted combination of latent calibration components learned from

related samples. This makes SGCP more expressive than discrete group-wise calibration, while still allowing samples with similar calibration behavior to share evidence through the components. Posterior averaging further stabilizes the resulting local score law (see Appendix B for details). Figure 2 illustrates how stochastic grouping induces a local score law for each sample.

The key question in Eq. (5) is therefore how to determine, for each sample x, which other samples should be regarded as 'related'. While there are many ways to define relevance around a point, such as finding similar points through feature embedding or clustering, recall that the ideal object for subgroup reliable calibration is the sample-specific score distribution $F_x^\star$ in Eq. (5). Accordingly, the relevant notion of locality is not simply geometric proximity in the input space, but similarity in calibration relevant score behavior. So we want to identify samples that are governed by similar score behavior. This motivates us to propose the latent calibration state and stochastic grouping map. Together, they determine how each sample should transfer calibration evidence from the latent calibration components. We next formalize these two components and the resulting local score law.

3.2 Learning a differentiable local score law

Latent calibration state. Let $h_\theta(x) \in \mathbb{R}^d$ denote the feature representation produced by the backbone classifier, and let $f_\theta(x) \in \Delta^{C-1}$ denote the corresponding predictive probability vector. To capture heterogeneity in score behavior beyond predefined groups, we associate each sample with a latent calibration state z. We model this state via a posterior distribution

$$q_\phi(z \mid x) = \mathcal{N}\Big(\mu_\phi\big(h_\theta(x)\big), \operatorname{diag}\big(\sigma_\phi^2(h_\theta(x))\big)\Big), \tag{6}$$

where $\mu_\phi(\cdot)$ and $\sigma_\phi(\cdot)$ are learned functions of the backbone representation. This stochastic formulation allows calibration behavior to remain ambiguous even when the observed feature vector is fixed. Samples with similar predictions may still display mixed or uncertain score behavior, and such ambiguity is naturally represented through posterior uncertainty in latent space.

Stochastic grouping map. For a latent draw $z \sim q_\phi(z \mid x)$, we obtain a simplex-valued membership vector $\tilde{w}(x,z) = \big(\tilde{w}_1(x,z), \ldots, \tilde{w}_K(x,z)\big) \in \Delta^{K-1}$ through a learnable differentiable mapping from the latent state to the probability simplex. A single latent draw represents only one possible realization of the posterior calibration state and therefore only one possible decomposition of the sample's local score law. What we seek, however, is a stable description of how the sample is associated with the latent calibration components. Accordingly, SGCP defines the *stochastic grouping map* by averaging memberships over the posterior distribution. Let $z^{(1)}, \ldots, z^{(M)} \overset{\text{i.i.d.}}{\sim} q_\phi(z \mid x)$. We define the posterior-averaged membership of sample x on the k-th latent calibration component

$$w_k(x) = \mathbb{E}_{z \sim q_\phi(\cdot \mid x)}[\tilde{w}_k(x,z)] \approx \frac{1}{M} \sum_{m=1}^{M} \tilde{w}_k\big(x, z^{(m)}\big), \qquad k = 1, \ldots, K. \tag{7}$$

This yields the stochastic grouping map $w(x) = \big(w_1(x), \ldots, w_K(x)\big) \in \Delta^{K-1}$, which determines how the latent calibration components are combined to form the learned local score law.

Aggregate local score law. Each latent calibration component $F_k : [0,1] \to [0,1]$ is a monotone mapping that represents one recurring calibration pattern in the score space. When needed, we write f_k for the corresponding density. Given the membership weight $w_k(x)$, SGCP defines the aggregate local score law at x as

$$\widehat{F}(s \mid x) = \sum_{k=1}^{K} w_k(x)\, F_k(s), \tag{8}$$

with corresponding local density $\widehat{f}(s \mid x) = \sum_{k=1}^{K} w_k(x)\, f_k(s)$. Equation (8) defines the central local score-law estimator in SGCP. It replaces the infeasible task of estimating an unconstrained $F_x^\star$ for every test point with a structured family generated by a small number of latent calibration components. In our implementation, each F_k is parameterized as a monotone differentiable mapping.

We next explain why local score-law normalization matters for subgroup reliability. If the local score law were known, the probability integral transform would map true-label nonconformity score onto a common percentile scale. The following proposition formalizes this property (proof in Appendix A.1).

Algorithm 1 Procedure for SGCP

Require: Training data $\mathcal{D}_{\mathrm{tra}}$, calibration data $\mathcal{D}_{\mathrm{cal}}$, test input X_{n+1}, nonconformity score $s(\cdot,\cdot)$, K latent calibration components, M posterior samples, hyperparameters λ, β, miscoverage level α

Ensure: Prediction set $\widehat{C}_\alpha(X_{n+1})$

1: Split $\mathcal{D}_{\mathrm{tra}}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$

2: **Fit SGCP:** on $\mathcal{D}_0$, learn f_θ, latent-state posterior $q_\phi(z \mid x)$ in Eq. (6), and stochastic grouping map $w(x)$ in Eq. (7); on $\mathcal{D}_1$, fit latent calibration components $\{F_k\}_{k=1}^K$ and local score law $\widehat{F}(\cdot \mid x)$ in Eq. (8).

3: **Calibration:** for each $(X_i, Y_i) \in \mathcal{D}_{\mathrm{cal}}$, compute $s_i = s(X_i, Y_i)$ and $V_i = \widehat{F}(s_i \mid X_i)$.

4: **Quantile:** sort $\{V_i : (X_i, Y_i) \in \mathcal{D}_{\mathrm{cal}}\}$ as $V_{(1)} \le \cdots \le V_{(m)}$, where $m = |\mathcal{D}_{\mathrm{cal}}|$, and set $\widehat{q}_{1-\alpha} = V_{(\lceil (m+1)(1-\alpha) \rceil)}$, with fallback $\widehat{q}_{1-\alpha} = 1$ if $\lceil (m+1)(1-\alpha) \rceil > m$.

5: **Test transform:** for each candidate label $y \in \mathcal{Y}$, compute $V_{n+1}^y = \widehat{F}\big(s(X_{n+1}, y) \mid X_{n+1}\big)$.

6: **Return:** $\widehat{C}_\alpha(X_{n+1}) = \big\{y \in \mathcal{Y} :\ V_{n+1}^y \le \widehat{q}_{1-\alpha}\big\}$.

Proposition 1. *Let $V(x,y) := \widehat{F}\big(s(x,y) \mid x\big)$ denote the transformed score. Suppose that the learned local score law coincides with the oracle local score law, namely, for P_X-almost every x, $\widehat{F}(t \mid x) = F_x^\star(t) := \mathbb{P}(S \le t \mid X = x)$, $\forall t \in [0,1]$. Assume furthermore that, for P_X-a.e. x, the conditional score distribution $S \mid X = x$ is continuous. Then, for any measurable subgroup $G \subseteq \mathcal{X}$ with $\mathbb{P}(X \in G) > 0$,*

$$V(X,Y) \mid \{X \in G\} \sim \mathrm{Unif}(0,1).$$

Proposition 1 explains why local score law learning can remove subgroup reliability inconsistency in the oracle case. In practice, SGCP learns this transform approximately, using the following objective.

Learning objective. We split the training data into $\mathcal{D}_{\mathrm{tra}} = \mathcal{D}_0 \cup \mathcal{D}_1$. The split $\mathcal{D}_0$ is used to learn the backbone predictor, the latent-state posterior, and the stochastic grouping map, while $\mathcal{D}_1$ is used to fit the latent calibration components and the learned local score law. An additional independent calibration set $\mathcal{D}_{\mathrm{cal}}$ is reserved for split conformal calibration. For each $(X_i, Y_i) \in \mathcal{D}_1$, we compute the true-label nonconformity score $s_i = s(X_i, Y_i)$, the membership weights $w_{ik} = w_k(X_i)$, and the component-wise transformed score $C_{ik} = F_k(s_i)$. Since F_k is monotone, C_{ik} can be viewed as the percentile position of s_i under the k-th latent calibration component.

The training objective encourages three complementary properties. First, within each component, the weighted distribution of transformed scores should be close to $\mathrm{Unif}(0,1)$. This yields the rank-uniformization loss $\mathcal{L}_{\mathrm{rank}}$. Second, the induced local mixture density $\widehat{f}(s_i \mid X_i) = \sum_{k=1}^K w_{ik} f_k(s_i)$ should explain the observed true-label scores, yielding the density-fitting loss $\mathcal{L}_{\mathrm{nll}}$. Finally, the stochastic grouping map should be informative: it should use the latent components diversely across the population while retaining sample-specific membership weights. We encourage this by maximizing the empirical mutual information between samples and a latent component variable Z, denoted by $\widehat{I}^{\,\mathrm{norm}}(X;Z)$, where the membership weights define $q_\phi(Z = k \mid X_i) = w_{ik}$.

Combining these terms, we optimize

$$\mathcal{L}_{\mathrm{SGCP}} = \mathcal{L}_{\mathrm{rank}} + \lambda \mathcal{L}_{\mathrm{nll}} - \beta \widehat{I}^{\,\mathrm{norm}}(X;Z), \tag{9}$$

where λ controls the strength of local density fitting and β controls the strength of stochastic grouping information maximization. Detailed parameterizations and explicit loss definitions are deferred to Appendix B. Algorithm 1 summarizes the overall procedure.

The following result, proved in Appendix A.2, shows that SGCP retains the standard finite-sample marginal validity guarantee of split conformal prediction, which contains the true label with probability at least $1-\alpha$. This guarantee does not depend on how accurately the local score law is learned, and the prediction set $\widehat{C}_\alpha(X_{n+1})$ remains valid.

Proposition 2. *Under the standard split-conformal exchangeability assumption, the prediction set returned by SGCP satisfies $\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\Big\} \ge 1-\alpha$.*

Marginal validity alone does not explain subgroup reliability. A further question is how the learned transform affects subgroup coverage. The following (proved in Appendix A.3) shows that subgroup coverage is controlled by the mismatch between subgroup and overall transformed-score distributions.

In the oracle property of Proposition 1, this mismatch vanishes for every subgroup. In practice, SGCP seeks to reduce it by learning a local score law transform through stochastic grouping. Appendix A.4 shows that uniformly approximating the oracle score law directly controls this mismatch.

Proposition 3. *For any measurable subgroup G with $\mathbb{P}(X \in G) > 0$, let Δ_G denote the Kolmogorov distance between the distribution of $V(X, Y)$ conditional on $X \in G$ and its marginal distribution. Under the assumptions of Proposition 2 and a continuous marginal distribution of $V(X, Y)$, subgroup coverage is lower bounded by $\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G\Big\} \geq 1 - \alpha - \Delta_G$.*

4 Numerical Experiments

Benchmarks and preprocessing. We evaluate SGCP on four benchmarks. For each dataset, let (X, Y, A) denote the feature vector, class label, and sensitive attributes, respectively. Below, we briefly describe how each dataset is processed and constructed for our evaluation.

(i) We generate synthetic (SYN) multiclass cardiopulmonary diagnosis benchmark with six disease categories, viz. Hypertension, Coronary Artery Disease, Heart Failure, Arrhythmia, Valvular Heart Disease, and Peripheral Artery Disease. Each sample contains four grouping attributes–gender, risk phenotype, age, and region–and six non-sensitive physiological variables. To mimic intersectional heterogeneity, individuals satisfying the XNOR relation between biological sex and risk phenotype, e.g., Female with Type I or Male with Type II, are assigned higher label uncertainty. This subgroup is therefore harder to classify and serves as the disadvantaged population. **(ii)** Nursery [21] is an admission-decision dataset for ranking nursery school applications, where family, socioeconomic, social, and health-related factors make it well suited for evaluating subgroup reliability. We use four priority ranks as targets and five categorical grouping attributes. Following [32, 30], we simulate controllable algorithmic bias by downsampling the subgroup defined by `parents` to 10%, adding label noise, and rounding labels to the nearest valid class. **(iii)** MIMIC-IV [13] is a real-world clinical tabular benchmark. We employ three binary grouping attributes, i.e., race, gender, and insurance status, for subgroup analysis, along with non-sensitive demographic and administrative covariates. We study two tasks: in-hospital mortality prediction and ICU-delay classification, where delays are grouped into short ($\leq$ 4 hours), moderate (4-12 hours), and prolonged ($>$ 12 hours) or no ICU admission. **(iv)** BACH [2] is a four-class histopathology benchmark for breast cancer diagnosis, viz. Normal, Benign, In Situ Carcinoma, and Invasive Car2cinoma. Since demographic attributes are unavailable, we use the class label as a proxy grouping variable, setting $A := Y$. Images are encoded by a visual backbone as $X_i = \Phi(I_i) \in \mathbb{R}^d$. Subgroup heterogeneity arises from visual variability across tissue and potential acquisition bias [17], so the class-conditional score laws $F_c(s) := \mathbb{P}(S \leq s \mid Y = c)$ may differ substantially across classes.

Across all datasets, we randomly split data into training, calibration, and test sets; for SGCP, the training set is further divided equally into $\mathcal{D}_0$ and $\mathcal{D}_1$. Sensitive attributes are used only for fairness evaluation and never as model inputs. Preprocessing details are given in Appendix D.1.

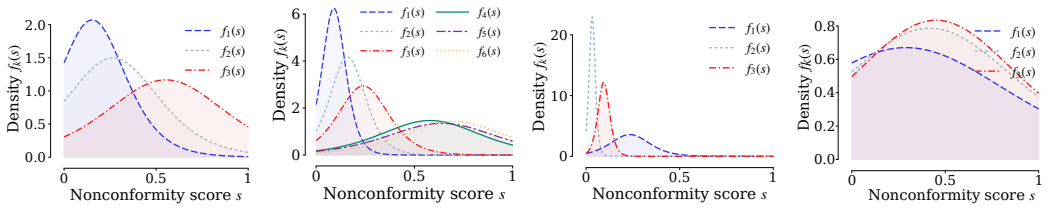
We benchmark SGCP against AFCP [32], Clustered CP [8], FaReG [30], RLCP [12], marginal CP, Partial Equalized [22], and Exhaustive Equalized. Partial Equalized protects one sensitive attribute at a time and unions the resulting sets, while Exhaustive Equalized protects all candidate attributes jointly as an explicit subgroup-equalization baseline. All methods use the same backbone and nonconformity score. Implementation details and hyperparameters are provided in Appendix D.2. See the Appendix E for additional experiments.

Evaluation criteria. We evaluate all methods in terms of overall validity, subgroup reliability, and efficiency. Let $\{(X'_i, Y'_i)\}_{i=1}^{N'}$ be the test set, and let $\mathcal{C}(X'_i)$ denote the prediction set at miscoverage level α. We report empirical *overall coverage* $\text{Cov.} = \frac{1}{N'} \sum_{i=1}^{N'} \mathbf{1}\{Y'_i \in \mathcal{C}(X'_i)\}$, which should be close to the target level $1 - \alpha$. Efficiency is measured by the *average set size* $\text{AvgSize} = \frac{1}{N'} \sum_{i=1}^{N'} |\mathcal{C}(X'_i)|$. Among methods with comparable coverage, smaller AvgSize is preferable.

Given a grouping attribute A and its induced groups, we are interested in coverage conditional on each group. For group $g \in \mathcal{G}(A)$, let $\mathcal{I}_g := \{i : A(X'_i) = g\}$. We define the estimated conditional coverage, hereafter *group coverage*, as $\hat{c}_g = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \mathbf{1}\{Y'_i \in \mathcal{C}(X'_i)\}$. We quantify subgroup disparity by the *coverage gap* (CovGap) $\text{CovGap}(A) = \frac{1}{|\mathcal{G}(A)|} \sum_{g \in \mathcal{G}(A)} |\hat{c}_g - (1 - \alpha)|$. Group

Table 1: Each method is run 20 times, and we report the mean and standard deviation. All results are reported for the largest sample size. Results with nominal coverage are shown in blue, optimal values are **boldfaced**, and AvgSize exceeding half of the total number of categories are highlighted in red.

Method	SYN			Nursery			MIMIC-IV			BACH		
	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓
Marginal	$89.52_{\pm 1.52}$	$2.309_{\pm 0.078}$	$14.5_{\pm 1.6}$	$84.15_{\pm 5.20}$	$2.170_{\pm 0.260}$	$20.4_{\pm 5.2}$	$90.79_{\pm 10.62}$	$1.226_{\pm 0.087}$	$27.1_{\pm 7.1}$	$89.20_{\pm 9.64}$	$2.894_{\pm 0.267}$	$15.7_{\pm 1.7}$
Partial	$94.50_{\pm 1.12}$	$2.419_{\pm 0.048}$	$6.2_{\pm 0.7}$	$88.97_{\pm 3.80}$	$1.850_{\pm 0.200}$	$12.1_{\pm 2.5}$	$94.50_{\pm 4.54}$	$1.258_{\pm 0.140}$	$8.0_{\pm 3.0}$	$96.57_{\pm 8.08}$	$3.385_{\pm 0.299}$	$7.6_{\pm 1.6}$
Exhaustive	$95.85_{\pm 0.91}$	$3.342_{\pm 0.053}$	$6.0_{\pm 0.6}$	$96.80_{\pm 3.10}$	$2.850_{\pm 0.220}$	$15.5_{\pm 3.2}$	$96.45_{\pm 4.22}$	$2.800_{\pm 0.142}$	$19.7_{\pm 3.2}$	$97.41_{\pm 6.89}$	$3.867_{\pm 0.118}$	$11.6_{\pm 1.5}$
RLCP	$93.10_{\pm 2.45}$	$2.540_{\pm 0.305}$	$4.5_{\pm 1.8}$	$88.67_{\pm 6.20}$	$1.220_{\pm 0.410}$	$13.5_{\pm 3.8}$	$92.50_{\pm 2.80}$	$1.350_{\pm 0.100}$	$1.7_{\pm 0.8}$	$91.88_{\pm 10.56}$	$2.311_{\pm 0.319}$	$8.1_{\pm 2.5}$
CluCP	$89.80_{\pm 1.78}$	$2.426_{\pm 0.050}$	$4.9_{\pm 0.8}$	$85.08_{\pm 1.60}$	$1.121_{\pm 0.064}$	$12.8_{\pm 1.3}$	$92.22_{\pm 1.28}$	$1.302_{\pm 0.059}$	$3.5_{\pm 0.8}$	$95.91_{\pm 3.59}$	$3.257_{\pm 0.220}$	$10.7_{\pm 2.6}$
AFCP	$91.77_{\pm 1.23}$	$2.432_{\pm 0.043}$	$4.6_{\pm 0.5}$	$92.85_{\pm 0.84}$	$1.085_{\pm 0.024}$	$14.1_{\pm 0.6}$	-	-	-	$95.00_{\pm 2.60}$	$2.830_{\pm 0.237}$	$6.3_{\pm 1.7}$
FaReG	$92.61_{\pm 0.98}$	$2.579_{\pm 0.029}$	$4.7_{\pm 0.5}$	$93.21_{\pm 0.62}$	$1.230_{\pm 0.012}$	$13.3_{\pm 0.6}$	$90.23_{\pm 0.09}$	$1.174_{\pm 0.001}$	$\mathbf{0.5}_{\pm 0.0}$	$93.60_{\pm 3.40}$	$2.959_{\pm 0.239}$	$7.4_{\pm 1.5}$
SGCP	$\mathbf{90.13}_{\pm 0.39}$	$\mathbf{2.225}_{\pm 0.035}$	$\mathbf{4.2}_{\pm 0.7}$	$\mathbf{91.65}_{\pm 0.70}$	$1.089_{\pm 0.035}$	$\mathbf{11.8}_{\pm 0.7}$	$\mathbf{90.13}_{\pm 0.07}$	$\mathbf{1.107}_{\pm 0.001}$	$0.8_{\pm 0.0}$	$\mathbf{91.90}_{\pm 2.70}$	$\mathbf{2.224}_{\pm 0.215}$	$\mathbf{6.2}_{\pm 1.1}$



(a) SYN (b) Nursery (c) MIMIC-IV (d) BACH

Figure 3: Learned latent component densities on four datasets. The components reveal distinct calibration regimes in the nonconformity-score space.

coverage closer to $1-\alpha$ and smaller CovGap indicate more reliable subgroup performance. All metrics other than AvgSize are reported as percentages.

4.1 Results and Discussion

Main results. Table 1 compares all methods across the four benchmarks (other experiments cf. Appendix E). AFCP is omitted on MIMIC-IV because its adaptive group selection requires an expensive repeated calibration procedure, making 20-run evaluation computationally prohibitive. We highlight AvgSize values exceeding half of the label space in red, as prediction sets of this size can reduce clinical utility, increase the cognitive burden on decision makers.

We highlight four main observations. **(i)** SGCP is the only method that consistently stays close to nominal coverage across all benchmarks while achieving strong subgroup reliability. Its empirical coverage remains near 0.9, and its CovGap is the smallest on SYN, Nursery, and BACH, with values 4.2, 11.8, and 6.2, respectively. On MIMIC-IV, it is close to the best result, with CovGap 0.8 compared with FaReG's 0.5. **(ii)** The subgroup reliability gains of SGCP are not obtained through indiscriminate conservatism. SGCP achieves the smallest or nearly smallest AvgSize across all datasets. Taken together, these results show that SGCP improves subgroup reliability without enlarging prediction sets. **(iii)** Existing baselines exhibit less favorable or less consistent tradeoffs. Marginal CP often leaves large subgroup disparities, while explicit subgroup-protection methods reduce disparity at the cost of inflated coverage and larger prediction sets. RLCP, CluCP, AFCP, and FaReG are competitive in some settings, but their gains are less uniform across benchmarks.

Figure 3 visualizes the latent component densities learned by SGCP. Across datasets, the components exhibit diverse shapes in the nonconformity score space. Some components concentrate near small scores, while others are broader or place more mass on larger scores. This suggests that SGCP learns nontrivial score law structure, rather than merely reproducing a single global score distribution. Importantly, these components should not be viewed as demographic groups or semantic subpopulations. They are latent score law building blocks. For each sample, SGCP mixes these shared components to form a sample dependent local score law. In this way, samples with similar calibration behavior can share evidence even when they do not belong to the same observed subgroup. This provides a mechanism for improving subgroup reliability without requiring calibration to be tied to a single observed group label.

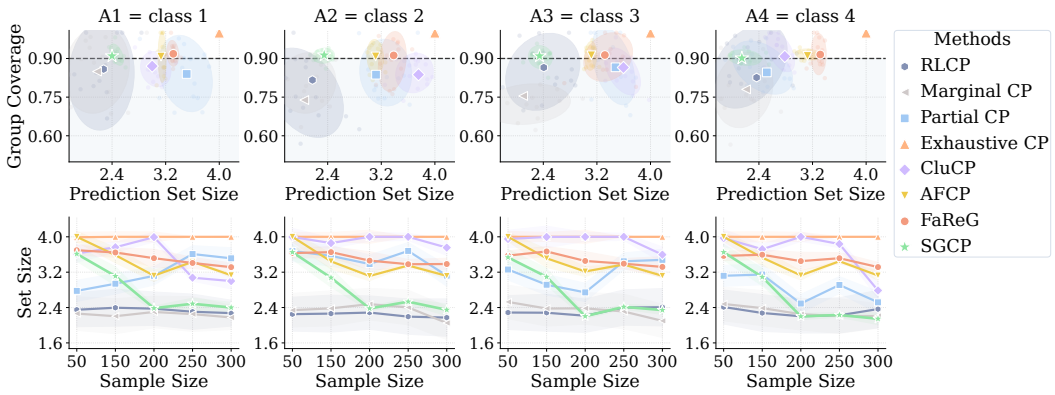


Figure 4: Group coverage on BACH dataset. SGCP remains close to the target coverage across classes while using smaller prediction sets than more conservative baselines.

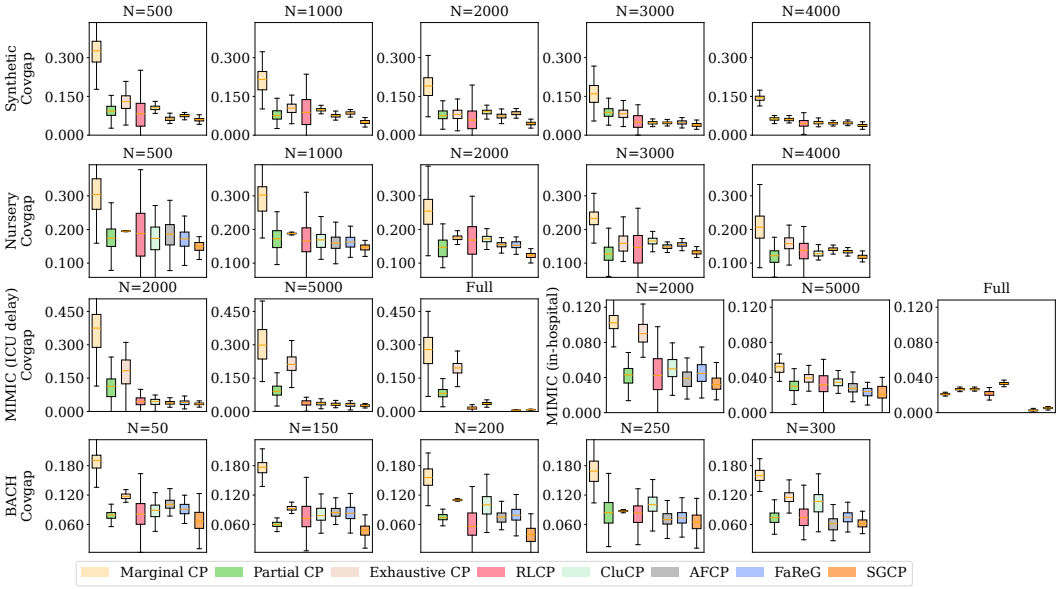


Figure 5: Distribution of CovGap over repeated runs under varying sample sparsity. **Takeaway:** SGCP maintains uniformly low CovGap across datasets and sparsity levels, indicating that its subgroup gains are stable rather than driven by a few favorable runs. By contrast, marginal and more conservative baselines often show either larger subgroup disparity or greater variability.

Group coverage. To better understand where the improvement comes from, we visualize subgroup coverage against average prediction-set size, together with how set size varies with subgroup sample size. Figure 4 corresponds to BACH, where the subgroup variable is the class label. Two patterns are clear. First, SGCP remains close to the nominal target coverage while lying on the more efficient side of the tradeoff: in the top row, its points stay near the horizontal reference line at 0.9 and are typically located to the left of competing methods. Second, in the bottom row, its prediction sets shrink steadily as subgroup sample size increases, indicating that it can use additional subgroup evidence efficiently while still adapting locally. Overall, these subgroup-level plots show that the gains of SGCP are realized across individual groups rather than only in aggregate.

Distribution of CovGap. Figure 5 shows the distribution of CovGap over repeated runs under varying levels of sample sparsity. Beyond low average disparity, SGCP also exhibits stability: across datasets and sample sizes, it is consistently among the methods with the lowest median CovGap and

relatively small variation across runs. This advantage is most pronounced in sparse-data settings, where subgroup calibration is most difficult. As sample size grows, the gap narrows for most methods, but SGCP remains competitive and typically retains one of the lowest CovGap values.

5 Conclusion

In this work, we discussed the limitation of standard split-conformal prediction in fairness-sensitive settings, where marginal validity can mask substantial subgroup miscoverage. We argued that subgroup reliability is closely tied to the local distribution of nonconformity scores, and that modeling this local score law provides a principled way to improve calibration across heterogeneous subpopulations. To this end, we proposed SGCP, a fairness-aware conformal prediction method that improves coverage for clinically meaningful subgroups without relying on predefined sensitive groups. Our theoretical analysis showed that SGCP preserves the standard split-conformal marginal coverage guarantee and accurate local score-law learning reduces subgroup coverage gaps by making calibration behavior more comparable across subpopulations. Extensive experiments across synthetic, and real-world benchmarks demonstrated that SGCP consistently improves subgroup reliability while maintaining efficient prediction sets. In future work, we plan to extend SGCP to personalized medicine, where reliable uncertainty quantification across diverse patient populations is crucial.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62572097). We gratefully acknowledge the support provided by the Center for HPC at the University of Electronic Science and Technology of China.

References

[1] A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

[2] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019.

[3] K. Bairaktari, J. Wu, and S. Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In *International Conference on Learning Representations (ICLR)*, 2025.

[4] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, 2023.

[5] M. Y. Cheung, A. Veeraraghavan, and G. Balakrishnan. COMPASS: Robust feature conformal prediction for medical segmentation metrics. In *International Conference on Learning Representations (ICLR)*, volume 2026, pages 134348–134381, 2026.

[6] S. Cortes-Gomez, C. M. Patiño, Y. Byun, S. Wu, E. Horvitz, and B. Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *International Conference on Learning Representations (ICLR)*, 2025.

[7] J. C. Cresswell, B. Kumar, Y. Sui, and M. Belbahri. Conformal prediction sets can cause disparate impact. In *International Conference on Learning Representations (ICLR)*, 2025.

[8] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64555–64576, 2023.

[9] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, and S. Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

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

[11] C. Gao, P. B. Gilbert, and L. Han. Bridging fairness and efficiency in conformal inference: a surrogate-assisted group-clustered approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

[12] R. Hore and R. F. Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2024.

[13] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV. *PhysioNet*, 2024. Version 3.1.

[14] A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. MIMIC-CXR Database. *PhysioNet*, Sept. 2019. Version 2.0.0.

[15] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[16] H. Kwon and D.-J. Kim. Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift. *Scientific Reports*, 16(1):10016, 2026.

[17] J. Lee and R. M. Nishikawa. Identifying women with mammographically- occult breast cancer leveraging gan-simulated mammograms. *IEEE Transactions on Medical Imaging*, 41(1):225–236, 2022.

[18] C. Lu, A. N. Angelopoulos, and S. Pomerantz. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, page 545–554, 2022.

[19] C. Lu, A. Lemay, K. Chang, K. Höbel, and J. Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12008–12016, 2022.

[20] N. Martinez, D. C. Patel, C. Reddy, G. Ganapavarapu, R. Vaculin, and J. Kalagnanam. Identifying homogeneous and interpretable groups for conformal prediction. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

[21] V. Rajkovic. Nursery. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5P88W.

[22] Y. Romano, R. F. Barber, C. Sabatti, and E. Candès. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2), 2020.

[23] Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 3581–3591, 2020.

[24] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[25] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, 2018.

[26] A. T. Vadlamani, A. Srinivasan, P. Maneriker, A. Payani, and S. Parthasarathy. A generic framework for conformal fairness. In *International Conference on Learning Representations (ICLR)*, 2025.

[27] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, pages 1–476. Springer Cham, 2005.

[28] V. Vovk, D. Lindsay, I. Nouretdinov, and A. Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

[29] Z. Wang and H. Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[30] S. Xu, Y. Zhou, Y. Tan, Z. Li, Y. Yao, T. Chen, F. Xu, and X. Ma. Fair conformal classification via learning representation-based groups. In *International Conference on Learning Representations (ICLR)*, 2026.

[31] D. Zhang, A. Chatzimparmpas, N. Kamali, and J. Hullman. Evaluating the utility of conformal prediction sets for ai-advised image labeling. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2024.

[32] Y. Zhou and M. Sesia. Conformal classification with equalized coverage for adaptively selected groups. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A Mathematical Proofs

A.1 Proof of Proposition 1

Proof. Let $S := s(x, y)$ and $V := \widehat{F}(S \mid X)$. Let λ denote the Lebesgue measure on $[0, 1]$. By the oracle assumption, for P_X-a.e. x,

$$\widehat{F}(\cdot \mid x) = F_x^\star(\cdot), \qquad F_x^\star(t) := \mathbb{P}(S \le t \mid X = x).$$

Moreover, by assumption, the conditional distribution of $S \mid X = x$ is continuous for P_X-a.e. x. Hence, by the probability integral transform, for P_X-a.e. x,

$$F_x^\star(S) \mid X = x \sim \mathrm{Unif}(0, 1).$$

Since $V = \widehat{F}(S \mid X) = F_X^\star(S)$ almost surely, it follows that, for every Borel set $A \subseteq [0, 1]$,

$$\mathbb{E}\big[\mathbf{1}_{\{V \in A\}} \mid X\big] = \lambda(A) \qquad \text{a.s.}$$

Now let $G \subseteq \mathcal{X}$ be any measurable subgroup with $\mathbb{P}(X \in G) > 0$. Using the tower property,

$$\begin{aligned}
\mathbb{P}(V \in A \mid X \in G) &= \frac{\mathbb{E}\big[\mathbf{1}_{\{V \in A\}} \mathbf{1}_{\{X \in G\}}\big]}{\mathbb{P}(X \in G)} \\
&= \frac{\mathbb{E}\big[\mathbf{1}_{\{X \in G\}} \mathbb{E}\big[\mathbf{1}_{\{V \in A\}} \mid X\big]\big]}{\mathbb{P}(X \in G)} \\
&= \frac{\mathbb{E}\big[\mathbf{1}_{\{X \in G\}} \lambda(A)\big]}{\mathbb{P}(X \in G)} \\
&= \lambda(A).
\end{aligned}$$

Therefore, for every set $A \subseteq [0, 1]$, $\mathbb{P}(V \in A \mid X \in G) = \lambda(A)$, which is precisely the statement that

$$V \mid \{X \in G\} \sim \mathrm{Unif}(0, 1).$$

□

Remark 1. *When this oracle local score law is recovered, the local percentile transform places scores from different inputs and subgroups on a common uniform scale, thereby eliminating subgroup reliability inconsistency. The objective of SGCP is to approximate this. Put differently, this proposition clarifies why we learn $\widehat{F}(s \mid x)$ instead of operating directly on the raw score s.*

A.2 Proof of Proposition 2

Throughout the proof, we condition on the training split $\mathcal{D}_{\mathrm{tra}}$. Thus the fitted transformation $(x, y) \mapsto \widehat{F}(s(x, y) \mid x)$ is fixed. For simplicity, we suppress this conditioning from the notation.

Let $m := |\mathcal{D}_{\mathrm{cal}}|$ and write $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i=1}^m$ Define the transformed calibration scores

$$V_i := \widehat{F}(s(X_i, Y_i) \mid X_i), \qquad i = 1, \dots, m,$$

and the transformed score of the test point under its true label

$$V_{\mathrm{test}} := \widehat{F}(s(X_{n+1}, Y_{n+1}) \mid X_{n+1}).$$

Let $V_{(1)} \le \cdots \le V_{(m)}$ denote the order statistics of the calibration scores, and set

$$k := \lceil (m + 1)(1 - \alpha) \rceil .$$

Define the split-conformal cutoff by $\widehat{q}_{1-\alpha} = V_{(k)}$, with the convention $\widehat{q}_{1-\alpha} = 1$ whenever $k > m$.

Proof. By assumption, conditional on $\mathcal{D}_{\mathrm{tra}}$, the calibration examples in $\mathcal{D}_{\mathrm{cal}}$ and the test point (X_{n+1}, Y_{n+1}) are exchangeable. Since the same deterministic measurable map is applied to each point, it follows that $V_1, \dots, V_m, V_{\mathrm{test}}$ are exchangeable conditional on $\mathcal{D}_{\mathrm{tra}}$.

Let $V_{(1)} \leq \cdots \leq V_{(m}$ denote the ordered transformed calibration scores, and define the split conformal cutoff by $\widehat{q}_{1-\alpha} = V_{(\lceil (m+1)(1-\alpha) \rceil)}$, with the convention $\widehat{q}_{1-\alpha} = 1$ whenever $\lceil (m+1)(1-\alpha) \rceil > m$. By exchangeability and the standard split conformal order-statistic argument, we have $\mathbb{P}\left(V_{\text{test}} \leq \widehat{q}_{1-\alpha} \mid \mathcal{D}_{\text{tra}}\right) \geq 1-\alpha$.

It remains to connect this event to the prediction set. By definition,

$$\widehat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ \widehat{F}\big(s(X_{n+1}, y) \mid X_{n+1}\big) \leq \widehat{q}_{1-\alpha} \right\}.$$

Substituting the true label Y_{n+1} gives

$$\left\{ Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \right\} = \left\{ \widehat{F}\big(s(X_{n+1}, Y_{n+1}) \mid X_{n+1}\big) \leq \widehat{q}_{1-\alpha} \right\} = \left\{ V_{\text{test}} \leq \widehat{q}_{1-\alpha} \right\}.$$

Hence,

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,\Big|\, \mathcal{D}_{\text{tra}} \Big) \geq 1-\alpha.$$

Taking expectations over $\mathcal{D}_{\text{tra}}$ gives

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \right\} = \mathbb{E}\left[\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,\middle|\, \mathcal{D}_{\text{tra}} \right\} \right] \geq 1-\alpha.$$

□

Remark 2. *In the current implementation, the stochastic grouping map $w(x)$ may be approximated by Monte Carlo averaging over latent posterior samples. The validity proof above continues to hold as long as the transformed score is computed using a rule that is fixed conditional on $\mathcal{D}_{\text{tra}}$, or, more generally, uses auxiliary randomness that is independent across calibration and test samples.*

A.3 Proof of Proposition 3

Proof. By definition of the prediction set,

$$\widehat{C}_\alpha(X_{n+1}) = \left\{ y \in \mathcal{Y} : \widehat{F}(s(X_{n+1}, y) \mid X_{n+1}) \leq \widehat{q}_{1-\alpha} \right\}.$$

Therefore,

$$\left\{ Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \right\} = \left\{ V_{\text{test}} \leq \widehat{q}_{1-\alpha} \right\}.$$

Now fix a measurable subgroup $G \subseteq \mathcal{X}$ with $\mathbb{P}(X \in G) > 0$. Since the calibration sample is independent of the test point, $\widehat{q}_{1-\alpha}$ is independent of (X_{n+1}, Y_{n+1}). Hence, conditional on $\widehat{q}_{1-\alpha}$,

$$\begin{aligned} &\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,\Big|\, X_{n+1} \in G, \widehat{q}_{1-\alpha} \Big) \\ &\quad = \mathbb{P}(V_{n+1} \leq \widehat{q}_{1-\alpha} \,|\, X_{n+1} \in G, \widehat{q}_{1-\alpha}) \\ &\quad = H_G(\widehat{q}_{1-\alpha}), \end{aligned}$$

where

$$H_G(t) := \mathbb{P}(V(X,Y) \leq t \mid X \in G).$$

Taking expectations over $\widehat{q}_{1-\alpha}$ gives

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,\Big|\, X_{n+1} \in G \Big) = \mathbb{E}[H_G(\widehat{q}_{1-\alpha})].$$

Next, let

$$H(t) := \mathbb{P}(V(X,Y) \leq t)$$

be the global CDF of the transformed score, and let $V^\dagger$ be an independent draw from this distribution. Since $V_1, \ldots, V_m, V^\dagger$ are exchangeable and H is continuous, the standard split-conformal rank argument gives

$$\mathbb{P}\big(V^\dagger \leq \widehat{q}_{1-\alpha}\big) = \frac{k}{m+1}.$$

On the other hand, conditioning on $\widehat{q}_{1-\alpha}$,

$$\mathbb{P}\big(V^\dagger \leq \widehat{q}_{1-\alpha} \,\big|\, \widehat{q}_{1-\alpha}\big) = H(\widehat{q}_{1-\alpha}).$$

Therefore,

$$\mathbb{E}[H(\widehat{q}_{1-\alpha})] = \frac{k}{m+1}.$$

Combining the two identities,

$$\begin{aligned}
&\left|\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,\Big|\, X_{n+1} \in G\Big) - \frac{k}{m+1}\right| \\
&\quad = |\mathbb{E}[H_G(\widehat{q}_{1-\alpha})] - \mathbb{E}[H(\widehat{q}_{1-\alpha})]| \\
&\quad = |\mathbb{E}[H_G(\widehat{q}_{1-\alpha}) - H(\widehat{q}_{1-\alpha})]| \\
&\quad \le \mathbb{E}[|H_G(\widehat{q}_{1-\alpha}) - H(\widehat{q}_{1-\alpha})|] \\
&\quad \le \sup_{t\in[0,1]} |H_G(t) - H(t)| = \Delta_G.
\end{aligned}$$

This proves the claimed subgroup coverage gap bound.

Finally, since

$$\frac{k}{m+1} = \frac{\lceil (m+1)(1-\alpha)\rceil}{m+1} \ge 1-\alpha,$$

we obtain

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,\Big|\, X_{n+1} \in G\Big) \ge 1-\alpha-\Delta_G.$$

□

Since $k/(m+1) \ge 1-\alpha$, this result implies that whenever a subgroup's transformed-score distribution is close to the overall transformed-score distribution, the global conformal cutoff keeps its coverage close to the nominal level $1-\alpha$.

Remark 3. *Proposition 3 bounds subgroup miscoverage in terms of the residual transformed-score mismatch Δ_G. This mismatch can be further related to local score-law estimation error. If the learned transform satisfies*

$$\sup_{x,s} |\widehat{F}(s \mid x) - F_x^\star(s)| \le \varepsilon,$$

then, by the probability integral transform,

$$\Delta_G \le 2\varepsilon$$

for every measurable subgroup G. Consequently,

$$\left|\mathbb{P}\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \mid X_{n+1} \in G\} - (1-\alpha)\right| \le 2\varepsilon + \frac{1}{m+1}.$$

Thus, accurate local score-law learning provides a sufficient condition for small subgroup coverage gaps.

A.4 Empirical Subgroup Coverage

We now relate the empirical subgroup coverage measured on an independent test sample to the population subgroup coverage above. Under the assumptions of Proposition 3, let $\mathcal{D}_G^{\text{test}} = \{(X_j^{(G)}, Y_j^{(G)})\}_{j=1}^{n_G}$ be an i.i.d. test sample drawn from the conditional distribution of (X,Y) given $X \in G$, independent of $\mathcal{D}_{\text{tra}}$ and $\mathcal{D}_{\text{cal}}$.

Corollary 1. *Let $m = |\mathcal{D}_{\text{cal}}|$ and $k = \lceil (m+1)(1-\alpha)\rceil$, and assume $k \le m$. Let $\widehat{\text{Cov}}_G$ be the coverage measured on an independent subgroup test sample of size n_G, denoted as*

$$\widehat{\text{Cov}}_G := \frac{1}{n_G}\sum_{j=1}^{n_G} \mathbf{1}\left\{Y_j^{(G)} \in \widehat{C}_\alpha(X_j^{(G)})\right\}.$$

Then, for any $\delta \in (0,1)$, conditional on $\mathcal{D}_{\text{tra}}$, with probability at least $1-\delta$ jointly over $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_G^{\text{test}}$,

$$\left|\widehat{\text{Cov}}_G - (1-\alpha)\right| \le \Delta_G + \sqrt{\frac{\log(4/\delta)}{2n_G}} + \sqrt{\frac{\log(4/\delta)}{2m}} + \left|\frac{k}{m} - (1-\alpha)\right|.$$

In particular,

$$\widehat{\mathrm{Cov}}_G \geq 1 - \alpha - \Delta_G - \sqrt{\frac{\log(4/\delta)}{2n_G}} - \sqrt{\frac{\log(4/\delta)}{2m}}.$$

Corollary 1 connects the theoretical subgroup discrepancy to the coverage measured on an independent finite subgroup test sample. When the learned transform reduces Δ_G and both the calibration sample size m and subgroup test sample size n_G are sufficiently large, the empirical subgroup coverage concentrates near the nominal level $1 - \alpha$. The two concentration terms account for uncertainty in the calibration cutoff and sampling variation in the subgroup test set, while the final deterministic term accounts for quantile discretization.

Proof. Throughout the proof, we condition on $\mathcal{D}_{\mathrm{tra}}$, so that the learned score transformation is fixed. Let $V_1, \ldots, V_m$ be the transformed calibration scores, let

$$\widehat{H}_m(t) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{V_i \leq t\},$$

and let $\widehat{q}_{1-\alpha} = V_{(k)}$. By continuity of the marginal transformed-score distribution, the calibration scores have no ties almost surely, and hence

$$\widehat{H}_m(\widehat{q}_{1-\alpha}) = \frac{k}{m}. \tag{A10}$$

Let H and H_G denote the marginal and subgroup CDFs of the transformed score, respectively, conditional on $\mathcal{D}_{\mathrm{tra}}$. By definition,

$$\Delta_G = \sup_{t \in [0,1]} |H_G(t) - H(t)|.$$

The Dvoretzky–Kiefer–Wolfowitz inequality gives

$$\mathbb{P}\left(\sup_{t \in [0,1]} \left| \widehat{H}_m(t) - H(t) \right| \geq \sqrt{\frac{\log(4/\delta)}{2m}} \right) \leq \frac{\delta}{2}. \tag{A11}$$

Conditional on $\mathcal{D}_{\mathrm{tra}}$ and $\mathcal{D}_{\mathrm{cal}}$, the prediction set $\widehat{C}_\alpha$ is fixed. Since the test points

$$(X_j^{(G)}, Y_j^{(G)}), \qquad j = 1, \ldots, n_G,$$

are i.i.d. draws from the conditional distribution of (X, Y) given $X \in G$, the indicators

$$Z_j := \mathbf{1}\left\{ Y_j^{(G)} \in \widehat{C}_\alpha(X_j^{(G)}) \right\}, \qquad j = 1, \ldots, n_G,$$

are i.i.d. Bernoulli random variables with mean

$$\mathrm{Cov}_G(\mathcal{D}_{\mathrm{cal}}) := \mathbb{P}\left\{ Y \in \widehat{C}_\alpha(X) \right\} = H_G(\widehat{q}_{1-\alpha}).$$

Therefore, by Hoeffding's inequality,

$$\mathbb{P}\left(\left| \widehat{\mathrm{Cov}}_G - \mathrm{Cov}_G(\mathcal{D}_{\mathrm{cal}}) \right| \geq \sqrt{\frac{\log(4/\delta)}{2n_G}} \right) \leq \frac{\delta}{2}. \tag{A12}$$

By Eqs. (A11) and (A12) and a union bound, both events hold jointly with probability at least $1 - \delta$. On this event,

$$\begin{aligned}
\left| \widehat{\mathrm{Cov}}_G - (1 - \alpha) \right| &\leq \left| \widehat{\mathrm{Cov}}_G - \mathrm{Cov}_G(\mathcal{D}_{\mathrm{cal}}) \right| + |H_G(\widehat{q}_{1-\alpha}) - H(\widehat{q}_{1-\alpha})| \\
&\quad + \left| H(\widehat{q}_{1-\alpha}) - \widehat{H}_m(\widehat{q}_{1-\alpha}) \right| + \left| \widehat{H}_m(\widehat{q}_{1-\alpha}) - (1 - \alpha) \right| \\
&\leq \sqrt{\frac{\log(4/\delta)}{2n_G}} + \Delta_G + \sqrt{\frac{\log(4/\delta)}{2m}} + \left| \frac{k}{m} - (1 - \alpha) \right|,
\end{aligned}$$

Algorithm 2 Detailed Training and Inference Procedure for SGCP

Require: Training data $\mathcal{D}_{\mathrm{tra}}$, calibration data $\mathcal{D}_{\mathrm{cal}}$, backbone classifier f_θ, nonconformity score $s(\cdot,\cdot)$, K latent calibration components, M posterior samples, hyperparameters λ, β, miscoverage level α

Ensure: Prediction set $\widehat{C}_\alpha(X_{N+1})$

Split $\mathcal{D}_{\mathrm{tra}}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$
Train the backbone predictor f_θ on $\mathcal{D}_0$
Initialize the stochastic assignment module and component laws $\{F_k\}_{k=1}^K$
for each optimization step on $\mathcal{D}_1$ **do**
    Sample a mini-batch $\mathcal{B} = \{(X_i, Y_i)\}_{i=1}^B$
    Compute backbone features, and nonconformity scores $s_i = s(X_i, Y_i)$
    Estimate posterior-averaged memberships from $q_\phi(z \mid X_i)$
    Compute component transforms $C_{ik} = F_k(S_i)$ and local densities $\widehat{f}(S_i \mid X_i) = \sum_{k=1}^K w_{ik} f_k(S_i)$
    Compute $\mathcal{L}_{\mathrm{rank}}$, $\mathcal{L}_{\mathrm{nll}}$, and $\widehat{I}_{\mathcal{B}}^{\mathrm{norm}}(X;Z)$
    Update assignment and component parameters by minimizing $\mathcal{L}_{SGCP}$
end for
Construct the learned local score transform $\widehat{F}(s \mid x) = \sum_{k=1}^K w_k(x) F_k(s)$
for each $(X_i, Y_i) \in \mathcal{D}_{\mathrm{cal}}$ **do**
    Compute $V_i = \widehat{F}(s(X_i, Y_i) \mid X_i)$
end for
Sort $\{V_i\}_{i=1}^m$ and compute the conformal cutoff $\widehat{q}_{1-\alpha} = V_{(\lceil (m+1)(1-\alpha) \rceil)}$ with fallback $\widehat{q}_{1-\alpha} = 1$ if $\lceil (m+1)(1-\alpha) \rceil > m$
for each candidate label $y \in \mathcal{Y}$ **do**
    Compute $V_{\mathrm{test}}^y = \widehat{F}(S(X_{n+1}, y) \mid X_{N+1})$
end for
return $\widehat{C}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : V_{n+1}^y \le \widehat{q}_{1-\alpha}\}$

where the last inequality uses Eq. (A10).

Finally, on the same high-probability event,

$$\begin{aligned}
\widehat{\mathrm{Cov}}_G &\ge \mathrm{Cov}_G(\mathcal{D}_{\mathrm{cal}}) - \sqrt{\frac{\log(4/\delta)}{2n_G}} \\
&\ge H(\widehat{q}_{1-\alpha}) - \Delta_G - \sqrt{\frac{\log(4/\delta)}{2n_G}} \\
&\ge \frac{k}{m} - \Delta_G - \sqrt{\frac{\log(4/\delta)}{2n_G}} - \sqrt{\frac{\log(4/\delta)}{2m}} \\
&\ge 1 - \alpha - \Delta_G - \sqrt{\frac{\log(4/\delta)}{2n_G}} - \sqrt{\frac{\log(4/\delta)}{2m}},
\end{aligned}$$

where the last inequality follows from $k/m \ge 1-\alpha$. □

B Model Details

This appendix specifies the concrete instantiation of the aggregate local score law in Eq. (8) and the corresponding training objective. The algorithm 2 demonstrates a specific process.

Latent calibration state. Let $h_\theta(x) \in \mathbb{R}^d$ denote the backbone feature representation and let $f_\theta(x) \in \Delta^{C-1}$ denote the corresponding predictive probability vector. We parameterize a Gaussian latent posterior $q_\phi(z \mid x) = \mathcal{N}\Big(\mu_\phi(h_\theta(x)), \mathrm{diag}(\sigma_\phi^2(h_\theta(x)))\Big)$.

Stochastic grouping map. Let $c_1, \ldots, c_K \in \mathbb{R}^r$ denote K learnable prototypes in the latent space. For a latent draw $z \sim q_\phi(z \mid x)$, define the assignment vector

$$\tilde{w}_k(x, z) = \frac{\exp(-\|z - c_k\|_2^2/T)}{\sum_{j=1}^K \exp(-\|z - c_j\|_2^2/T)}, \qquad k = 1, \ldots, K,$$

where $T > 0$ is a temperature parameter. Given $z^{(1)}, \ldots, z^{(M)} \overset{\text{i.i.d.}}{\sim} q_\phi(z \mid x)$, the posterior-averaged membership vector $w(x) = (w_1(x), \ldots, w_K(x)) \in \Delta^{K-1}$ is estimated by

$$w_k(x) = \mathbb{E}_{z \sim q_\phi(\cdot|x)}[\tilde{w}_k(x, z)] \approx \frac{1}{M} \sum_{m=1}^M \tilde{w}_k(x, z^{(m)}), \qquad k = 1, \ldots, K.$$

Remark 4. *Marginal conformal validity only requires global exchangeability, so it can hold even when score behavior varies substantially across subgroups. Such heterogeneity can make a single global calibration rule locally unreliable. SGCP addresses this by learning latent calibration components that shift calibration evidence among samples with similar score behavior, providing a practical way to approximate oracle local score laws without assuming exact local exchangeability.*

Remark 5. *In SGCP, posterior sampling approximates a stable stochastic grouping map rather than adding randomness to conformal calibration. For fixed x, draw $z^{(1)}, \ldots, z^{(M)} \overset{\text{i.i.d.}}{\sim} q_\phi(z \mid x)$ and set*

$$\widehat{w}_k^{(M)}(x) = \frac{1}{M} \sum_{m=1}^M \tilde{w}_k(x, z^{(m)}), \qquad w_k(x) = \mathbb{E}_{z \sim q_\phi(\cdot|x)}[\tilde{w}_k(x, z)].$$

The induced local score law is $\widehat{F}^{(M)}(s \mid x) = \sum_{k=1}^K \widehat{w}_k^{(M)}(x) F_k(s)$.

Samples share calibration evidence through the common components $\{F_k\}_{k=1}^K$, and posterior averaging reduces Monte Carlo variability in the sample-specific weights. As $M \to \infty$, the law of large numbers gives $\widehat{w}_k^{(M)}(x) \to w_k(x)$ almost surely, and therefore

$$\widehat{F}^{(M)}(s \mid x) \to \widehat{F}(s \mid x) = \sum_{k=1}^K w_k(x) F_k(s) \quad \textit{a.s.}$$

B.1 Interpretation and Estimation of Latent Calibration Components

What do the components represent? The latent calibration components in SGCP describe recurring patterns in the distribution of true-label nonconformity scores. Let $S = s(X, Y)$ and $F_x^*(t) = \mathbb{P}(S \le t \mid X = x)$. Ideally, this conditional CDF would allow each sample to interpret its nonconformity score relative to its own score distribution. Since estimating F_x^* separately for every x is statistically difficult, SGCP approximates it through a mixture of shared components,

$$\widehat{F}(t \mid x) = \sum_{k=1}^K w_k(x) F_k(t). \tag{A13}$$

Each F_k is a CDF that captures a recurring pattern of score behavior, including its location, dispersion, and tail characteristics. It is not a cluster center or a component-specific conformal threshold. The weights $w_k(x)$ determine how these shared score distributions contribute to the local score law of sample x. Thus, the components describe what calibration behavior can be reused, while the weights determine how that behavior is combined for an individual sample. A sample may draw information from several components without being assigned to a single latent group.

How are the components learned? SGCP learns these components through calibration-relevant score behavior. The backbone representation is mapped to a Gaussian latent state $q_\phi(z \mid x)$, and each latent draw is softly assigned to learned prototypes. Averaging these assignments over the latent posterior gives the stochastic grouping weights $w_k(x)$.

The three terms in the learning objective serve complementary purposes. The rank-uniformization loss encourages component CDFs to map their associated scores to approximately uniform percentiles. The local likelihood loss encourages the resulting mixture to explain the observed true-label scores.

The grouping regularizer discourages collapsed or uniformly diffuse assignments, allowing different components to capture distinct score patterns. Posterior averaging further stabilizes the routing weights while retaining uncertainty over component membership.

After learning the routing mechanism, the final component CDFs are estimated on the held-out training-side split $\mathcal{D}_1$. Specifically, for $s_i = s(X_i, Y_i)$, the responsibility-weighted empirical CDF of component k is

$$\widehat{F}_k(t) = \frac{\sum_{i \in \mathcal{D}_1} w_k(X_i) \mathbb{I}\{s_i \leq t\}}{\sum_{i \in \mathcal{D}_1} w_k(X_i)}. \tag{A14}$$

Each training score therefore contributes fractionally to the component distributions according to its membership weights. The resulting local score law combines these empirical CDFs using the weights of the target sample.

Connection to subgroup reliability. For a candidate label y, SGCP transforms the raw score into $V(x, y) = \widehat{F}(s(x, y) \mid x)$. If the estimated local score law coincides with F_x^* and the conditional score distribution is continuous, the true-label transformed scores are uniformly distributed within every measurable subgroup, as established in Proposition 1. In practice, learning shared score distributions seeks to bring the transformed-score distributions of different subgroups closer together. This allows a single pooled conformal threshold to have a more comparable interpretation across heterogeneous samples, without requiring predefined sensitive groups or separate subgroup calibration sets. The independent conformal calibration step retains finite-sample marginal coverage, while subgroup reliability depends on how well the learned local score laws approximate the underlying score distributions.

B.2 Interpretability and Component Auditing

Interpretation at the prediction level. SGCP provides a statistical interpretation of how shared score distributions contribute to an individual prediction. For an input x and a candidate label y, the transformed score admits the decomposition

$$V(x, y) = \widehat{F}(s(x, y) \mid x) = \sum_{k=1}^{K} w_k(x) F_k(s(x, y)). \tag{A15}$$

An auditor can inspect the raw score $s(x, y)$, the routing weights $w_k(x)$, the component percentiles $F_k(s(x, y))$, and their weighted contributions. The candidate label is retained when $V(x, y) \leq \widehat{q}_{1-\alpha}$. Since the weights sum to one, this decision can also be examined through

$$\widehat{q}_{1-\alpha} - V(x, y) = \sum_{k=1}^{K} w_k(x) \left[\widehat{q}_{1-\alpha} - F_k(s(x, y))\right]. \tag{A16}$$

A positive term supports inclusion of the candidate label, whereas a negative term opposes inclusion; their sum determines the final decision. This interpretation concerns the calibration calculation for a fitted model and a fixed cutoff. It makes the role of each score regime directly inspectable without requiring the components to correspond to named clinical groups.

Component structure under temporal transfer. We conducted a post hoc audit using five runs of the MIMIC-IV temporal transfer experiment. Component indices are arbitrary across runs, so we aligned them by Hungarian matching on a fixed diagnostic cohort of 2,000 patients. Table A2 summarizes the aligned components.

The three target components exhibit progressively higher mean nonconformity scores, prediction error rates, and average set sizes, consistent with increasingly difficult calibration behavior. Their relative prevalence also changes substantially between cohorts. The mass of the component with the highest mean score increases from 33.05% in the source cohort to 57.11% in the target cohort, while that of the component with the lowest mean score decreases from 33.71% to 15.19%. These changes provide a summary of how the target population is represented by the learned score regimes.

Associations with observed attributes. We also examined whether the components have observable associations with patient characteristics. The proportion of patients with public insurance

Table A2: Post hoc component audit for the MIMIC-IV temporal transfer experiment using five runs. Predictive statistics describe the aligned target components; component masses are reported for both cohorts.

Statistic	Component 1	Component 2	Component 3
Responsibility-weighted mean score	0.058	0.158	0.394
Top-1 error rate (%)	2.35	6.82	25.86
Average prediction set size	1.033	1.162	1.582
Source component mass (%)	33.71	33.25	33.05
Target component mass (%)	15.19	27.70	57.11

increases from 33.70% to 52.41% and 65.90% across the three aligned components. Such associations help characterize the composition of the learned score regimes and can guide further investigation of where calibration difficulty is concentrated. They do not establish that the components are defined by insurance status or that insurance causes the observed differences in score behavior. The interpretation supported here is statistical and operational; clinical and causal meanings remain to be established.

B.3 Training Objective

We train the stochastic assignment module and the component parameters on $\mathcal{D}_1$, while keeping the backbone fixed. For each $(X_i, Y_i) \in \mathcal{D}_1$, define

$$S_i := s(X_i, Y_i), \qquad w_{ik} := w_k(X_i), \qquad C_{ik} := F_k(S_i).$$

The latent components are intended to serve as reusable local score laws. Thus, scores routed to the same component should be comparable after being mapped to their component-wise percentile scale, while the induced local mixture density should still explain the observed true-label nonconformity scores. These two requirements lead to the rank-uniformization and local density-fitting terms below.

Rank-uniformization. For a mini-batch $\mathcal{B}$ of size B and a fixed component k, let $\{C_{(1)k}, \ldots, C_{(B)k}\}$ be the sorted values of $\{C_{ik} : i \in \mathcal{B}\}$, and let $w_{(1)k}, \ldots, w_{(B)k}$ denote the correspondingly reordered weights. Normalize the weights within component k by

$$\pi_{(b)k} = \frac{w_{(b)k}}{\sum_{\ell=1}^{B} w_{(\ell)k}}.$$

The weighted rank midpoint is $r_{(b)k} = \sum_{\ell<b} \pi_{(\ell)k} + \frac{1}{2}\pi_{(b)k}$. We define the component-wise rank-uniformization loss as

$$\ell_{\mathrm{rank}}^{(k)} = \sum_{b=1}^{B} \pi_{(b)k} \left(C_{(b)k} - r_{(b)k}\right)^2.$$

Averaging over active components gives

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathcal{K}_{\mathrm{act}}|} \sum_{k \in \mathcal{K}_{\mathrm{act}}} \ell_{\mathrm{rank}}^{(k)}, \qquad \mathcal{K}_{\mathrm{act}} = \left\{ k : \sum_{i \in \mathcal{B}} w_{ik} > \tau \right\}.$$

Local density fitting. Let f_k denote the density associated with F_k. The local mixture density at (X_i, S_i) is $\widehat{f}(S_i \mid X_i) = \sum_{k=1}^{K} w_{ik} f_k(S_i)$. We fit the observed true-label scores by the negative log-likelihood

$$\mathcal{L}_{\mathrm{nll}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \widehat{f}(S_i \mid X_i).$$

The rank term makes the component transforms behave like percentile maps, whereas the likelihood term ties the learned score laws to the empirical score distribution.

Information maximization for stochastic grouping. The preceding two terms specify how the component score laws should behave, but they do not by themselves ensure that the grouping map uses the components in an informative way. For example, a balance constraint on the average component mass can prevent all samples from collapsing onto a single component, but it still permits the opposite degenerate solution in which every sample is assigned nearly uniformly to all components. Such assignments are balanced in aggregate, yet they carry little sample-specific information and tend to make the learned component laws indistinguishable.

We therefore encourage the component assignment to retain information about the input. For a mini-batch $\mathcal{B}$, view the assignment as a conditional categorical distribution $q_\phi(Z = k \mid X_i) = w_{ik}$, where Z denotes the latent component index. The induced marginal component distribution on the mini-batch is

$$\bar{w}_k = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_{ik}, \qquad \bar{w} = (\bar{w}_1, \ldots, \bar{w}_K).$$

The empirical mutual information between inputs and component assignments is

$$\widehat{I}_{\mathcal{B}}(X; Z) = H(\bar{w}) - \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} H(w_i),$$

where

$$H(\bar{w}) = -\sum_{k=1}^{K} \bar{w}_k \log \bar{w}_k, \qquad H(w_i) = -\sum_{k=1}^{K} w_{ik} \log w_{ik}.$$

We use the normalized form

$$\widehat{I}_{\mathcal{B}}^{\mathrm{norm}}(X; Z) = \frac{1}{\log K} \left[H(\bar{w}) - \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} H(w_i) \right].$$

The marginal entropy term encourages all components to be used at the batch level, while the conditional entropy term encourages each sample to have a sharper, more informative assignment. Thus, maximizing $\widehat{I}_{\mathcal{B}}^{\mathrm{norm}}(X; Z)$ discourages both hard component collapse and soft uniform assignment.

Full objective. Taking the rank-uniformization loss as the reference scale, we optimize

$$\mathcal{L}_{\mathrm{SGCP}} = \mathcal{L}_{\mathrm{rank}} + \lambda \mathcal{L}_{\mathrm{nll}} - \beta \widehat{I}_{\mathcal{B}}^{\mathrm{norm}}(X; Z),$$

where λ controls the strength of local density fitting and β controls the strength of stochastic-grouping information maximization.

C Complexity Analysis

Beyond backbone training, SGCP adds three sources of computation: stochastic grouping, local score-law fitting, and conformal calibration on transformed scores. Let N denote the number of training samples used for learning the grouping map and local score laws, N_{cal} the calibration size, and N_{test} the number of test samples. Treating the number of epochs, posterior samples, latent components, latent dimension, classes, and batch size as constants, the training overhead is linear in N, up to an optional $\mathcal{O}(N \log N)$ sorting step for constructing empirical component CDFs.

Calibration requires computing one transformed score per calibration sample and selecting a conformal quantile, costing $\mathcal{O}(N_{\mathrm{cal}} \log N_{\mathrm{cal}})$ with sorting or $\mathcal{O}(N_{\mathrm{cal}})$ with linear-time selection. At test time, the grouping map is evaluated once per input and transformed scores are computed for all candidate labels, giving constant overhead per test sample and total cost $\mathcal{O}(N_{\mathrm{test}})$.

Overall, under the standard mini-batch implementation, the additional complexity beyond backbone training is $\mathcal{O}(N + N_{\mathrm{cal}} \log N_{\mathrm{cal}} + N_{\mathrm{test}})$, or $\mathcal{O}(N + N_{\mathrm{cal}} + N_{\mathrm{test}})$ with linear-time quantile selection.

D Detailed Experimental Setup

D.1 Additional Benchmark Details

In this section, we describe the dataset preprocessing in detail. A summary is provided in Table A3.

Table A3: Summary of datasets and grouping attributes used for evaluation.

Dataset	Modality	#Classes	#Attrs	Sensitive / grouping attributes
SYN	tabular	6	3	A_1 = gender, A_2 = phenotype, A_3 = age
Nursery	tabular	4	5	A_1 = parents, A_2 = children, A_3 = finance, A_4 = social, A_5 = health
MIMIC-IV	tabular	2/3	3	A_1 = minority, A_2 = gender, A_3 = public insurance
BACH	image	4	-	class label (for class-wise reliability analysis)

Synthetic. The synthetic benchmark is a controlled multiclass cardiopulmonary diagnosis task with six labels, $Y \in \{1, \ldots, 6\}$, corresponding to Hypertension, Coronary Artery Disease, Heart Failure, Arrhythmia, Valvular Heart Disease, and Peripheral Artery Disease.

Each sample contains four grouping attributes, $A = (A_{\text{sex}}, A_{\text{pheno}}, A_{\text{age}}, A_{\text{region}})$, and six non-sensitive physiological variables $Z = (Z_1, \ldots, Z_6)$, where $Z_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1)$. We draw

$$A_{\text{sex}} \sim \text{Bernoulli}(1/2), \qquad A_{\text{pheno}} \sim \text{Bernoulli}(1/2),$$

with $A_{\text{sex}} \in \{0, 1\}$ denoting Female/Male and $A_{\text{pheno}} \in \{0, 1\}$ denoting Type I/Type II. Age is assigned cyclically over $\{\text{Pediatric}, \text{Adult}, \text{Geriatric}\}$, and region is sampled uniformly from $\{\text{North}, \text{South}, \text{East}, \text{West}\}$.

To induce intersection-dependent difficulty, we define the disadvantaged subgroup through the XNOR relation between sex and phenotype,

$$G_{\text{bias}} = \{i : \ A_{\text{sex},i} = A_{\text{pheno},i}\},$$

which includes Female–Type I and Male–Type II individuals.

Labels are generated from a latent multiclass mechanism based on $X_i = (A_i, Z_i)$. We first construct a class-probability vector $\pi(X_i) \in \Delta^5$ over the six disease classes. Outside ambiguous regions of the feature space, labels follow the dominant class implied by $\pi(X_i)$. Inside ambiguous regions, we introduce subgroup-dependent uncertainty by mixing $\pi(X_i)$ with the uniform distribution $u = (1/6, \ldots, 1/6)$:

$$Y_i \sim \begin{cases} \text{Cat}\big((1-\delta_1)\pi(X_i) + \delta_1 u\big), & i \in G_{\text{bias}}, \\ \text{Cat}\big((1-\delta_0)\pi(X_i) + \delta_0 u\big), & i \notin G_{\text{bias}}, \end{cases} \qquad \delta_1 > \delta_0.$$

The parameters δ_1 and δ_0 control label ambiguity for the disadvantaged subgroup and the remaining population, respectively. Thus, G_{bias} exhibits systematically higher conditional uncertainty while the global task structure remains unchanged, yielding a controlled setting for studying subgroup-dependent calibration difficulty.

Nursery. Nursery is a public multiclass tabular benchmark derived from a hierarchical decision problem for ranking nursery-school applications. Following standard preprocessing, we remove the `recommend` class, which contains only two samples, and encode the remaining labels into four admission-priority levels, $Y \in \{0, 1, 2, 3\}$.

Each instance contains eight categorical covariates describing household background and application context, including parents' occupation, number of children, housing, financial standing, social conditions, and health status. All categorical variables are label-encoded before training. We use five covariates as candidate sensitive attributes,

$$A = (A_{\text{parents}}, A_{\text{children}}, A_{\text{finance}}, A_{\text{social}}, A_{\text{health}}),$$

corresponding to parents' occupation, number of children, financial standing, social conditions, and health status. These attributes are used only for subgroup construction and fairness evaluation.

To induce controlled subgroup disadvantage, we follow the intervention protocol of prior work [32, 30]. We define

$$G_{\text{nur}} = \{i : X_{\text{parents},i} = 1\},$$

corresponding to a particular preprocessed category of the `parents` attribute. We then modify only this subgroup by downsampling it to 10% of its original size and injecting label uncertainty. Specifically, labels are perturbed by independent noise $\varepsilon_i \sim \mathcal{U}(-4, 4)$, then rounded and clipped to $\{0, 1, 2, 3\}$. This creates a subgroup that is both underrepresented and noisier than the rest of the population, serving as the disadvantaged group in our fairness evaluation.

Table A4: Dataset specific hyperparameter settings for SGCP.

Hyperparameter	SYN	Nursery	M-ICU	M-Hosp	BACH	Description
(i) Latent calibration components						
K	3	6	3	3	6	number of latent calibration components
d_z	16	8	8	8	64	latent calibration dimension
hidden dim	64	64	32	32	64	hidden dimension of grouping network
(ii) Score law learning						
SGCP train epochs	100	100	100	100	100	number of SGCP training epochs
SGCP learning rate	10^{-3}	10^{-2}	10^{-3}	10^{-3}	10^{-3}	learning rate for SGCP training
λ	0.5	0.05	10^{-3}	10^{-3}	0.5	weight of negative log likelihood loss
β	0.1	10^{-3}	0.5	10^{-3}	5×10^{-5}	weight of balance regularization
M_{train}	3	6	3	3	8	posterior samples during training
M_{eval}	8	8	8	16	8	posterior samples during evaluation
(iii) Backbone & optimization						
train epochs	200	200	40	40	100	number of backbone training epochs
learning rate	5×10^{-4}	10^{-3}	5×10^{-4}	10^{-3}	10^{-3}	learning rate for backbone training
batch size	32	128	128	128	32	mini batch size
weight decay	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	optimizer weight decay

Note: M-ICU denotes MIMIC-IV (ICU delay) and M-Hosp denotes MIMIC-IV (in hospital).

Table A5: Ablation study of SGCP over 20 runs. Results with nominal coverage are shown in blue, optimal values are **boldfaced**, and AvgSize exceeding half of the total number of categories is highlighted in red. CovGap is reported in percent.

Method	**SYN**			**Nursery**			**MIMIC**			**BACH**		
	Cov.	AvgSize	CovGap	Cov.	AvgSize	CovGap	Cov.	AvgSize	CovGap	Cov.	AvgSize	CovGap
Default SGCP	**90.13**$_{\pm 0.39}$	**2.225**$_{\pm 0.035}$	**4.2**$_{\pm 0.7}$	**91.65**$_{\pm 0.70}$	1.089$_{\pm 0.035}$	**11.8**$_{\pm 0.7}$	**90.13**$_{\pm 0.07}$	**1.107**$_{\pm 0.001}$	0.8$_{\pm 0.0}$	**91.90**$_{\pm 2.70}$	**2.224**$_{\pm 0.215}$	**6.2**$_{\pm 1.1}$
(a) w/o Stochastic	89.82$_{\pm 1.48}$	2.348$_{\pm 0.081}$	7.35$_{\pm 1.18}$	91.38$_{\pm 1.15}$	1.123$_{\pm 0.051}$	13.42$_{\pm 1.76}$	89.94$_{\pm 0.96}$	1.053$_{\pm 0.013}$	2.36$_{\pm 0.48}$	91.76$_{\pm 4.42}$	2.447$_{\pm 0.308}$	8.84$_{\pm 2.35}$
(b) Feature space	90.96$_{\pm 1.96}$	2.476$_{\pm 0.127}$	7.18$_{\pm 1.46}$	90.92$_{\pm 1.43}$	1.248$_{\pm 0.079}$	14.16$_{\pm 2.05}$	90.48$_{\pm 1.47}$	1.179$_{\pm 0.044}$	4.08$_{\pm 1.08}$	89.61$_{\pm 5.74}$	3.094$_{\pm 0.446}$	12.41$_{\pm 3.16}$
(c) w/o rank	89.21$_{\pm 4.42}$	1.898$_{\pm 0.246}$	8.42$_{\pm 4.10}$	88.63$_{\pm 2.94}$	1.302$_{\pm 0.118}$	15.31$_{\pm 2.41}$	90.03$_{\pm 5.41}$	1.334$_{\pm 0.146}$	5.49$_{\pm 3.71}$	91.18$_{\pm 6.31}$	3.491$_{\pm 0.541}$	9.26$_{\pm 3.02}$
(d) w/o NLL	88.36$_{\pm 1.17}$	2.074$_{\pm 0.108}$	5.39$_{\pm 1.08}$	92.39$_{\pm 0.92}$	1.446$_{\pm 0.094}$	12.43$_{\pm 1.19}$	90.02$_{\pm 2.96}$	1.351$_{\pm 0.078}$	2.96$_{\pm 0.77}$	93.11$_{\pm 2.96}$	2.846$_{\pm 0.317}$	7.44$_{\pm 1.47}$
(e) w/o MI	92.28$_{\pm 3.41}$	2.438$_{\pm 0.241}$	6.53$_{\pm 2.72}$	91.26$_{\pm 2.43}$	1.149$_{\pm 0.147}$	14.72$_{\pm 1.42}$	89.84$_{\pm 6.31}$	1.047$_{\pm 0.116}$	4.95$_{\pm 2.31}$	90.91$_{\pm 7.82}$	2.347$_{\pm 0.472}$	9.31$_{\pm 2.11}$

MIMIC-IV. We construct our clinical benchmark from MIMIC-IV at the hospital-admission level. Each admission is linked with ICU-stay information and diagnosis-code counts, yielding tabular covariates from demographic, administrative, and utilization records. Age is computed from `anchor_age`, adjusted by the difference between admission year and anchor year, and clipped to $[0, 89]$. We also derive ICU-related features, including ICU length of stay, time from hospital admission to first ICU entry, and the number of diagnosis codes.

For subgroup evaluation, we define three binary grouping attributes. The race indicator `minority` equals 1 if the recorded race string does not contain `WHITE`; `gender_m` equals 1 for male patients; and `public_insurance` equals 1 if the primary insurance is `MEDICARE` or `MEDICAID`. These attributes are used only for subgroup construction and fairness evaluation. Model inputs include age, one-hot indicators for insurance type, admission type, and marital status, ICU length of stay, and the number of diagnosis codes; ICU length of stay is excluded for ICU-delay prediction.

We consider two tasks. The first is in-hospital mortality prediction, with binary labels derived from `hospital_expire_flag`. Results for this task are reported as MIMIC-IV (in-hospital) in Figures 5, A9, and A10. The second is ICU-delay prediction, reported as MIMIC-IV (ICU delay), where admissions are categorized by the time from hospital admission to first ICU entry:

$$Y = \begin{cases} 0, & \text{if ICU delay } \leq 4 \text{ hours}, \\ 1, & \text{if ICU delay } \in (4, 12] \text{ hours}, \\ 2, & \text{if ICU delay } > 12 \text{ hours or no ICU admission occurs}. \end{cases}$$

Together, these tasks provide a real-world clinical benchmark for evaluating fairness across race, gender, and insurance-defined subgroups.

BACH. The Breast Cancer Histology (BACH) dataset [2] is a histopathology benchmark for evaluating subgroup reliability under visual heterogeneity. We use Part A, which contains 400 H&E-stained microscopy images evenly distributed across four classes: normal, benign, in situ carcinoma, and invasive carcinoma. Images were acquired at $20\times$ magnification with resolution

Table A6: Performance comparison w.r.t. sample sparsity on the Synthetic Dataset (Target $1-\alpha = 0.90$). **Takeaway:** SGCP remains consistently close to nominal coverage across all sample sizes while achieving the smallest subgroup coverage gap and comparatively small prediction sets. And the performance of each specific sensitive group is shown in the Fig. A6.

Method	N = 500			N = 1000			N = 2000			N = 4000		
	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓
Marginal	$90.15_{\pm 6.50}$	$2.150_{\pm 0.420}$	$32.5_{\pm 6.5}$	$90.20_{\pm 4.20}$	$1.950_{\pm 0.400}$	$21.5_{\pm 6.2}$	$90.12_{\pm 4.10}$	$1.820_{\pm 0.380}$	$19.7_{\pm 6.0}$	$89.52_{\pm 1.52}$	$2.309_{\pm 0.078}$	$14.5_{\pm 1.6}$
Partial	$95.50_{\pm 3.20}$	$4.520_{\pm 0.300}$	$9.5_{\pm 2.8}$	$94.20_{\pm 3.00}$	$2.950_{\pm 0.280}$	$7.9_{\pm 2.5}$	$93.50_{\pm 2.80}$	$2.820_{\pm 0.250}$	$7.5_{\pm 2.2}$	$94.50_{\pm 1.12}$	$2.419_{\pm 0.048}$	$6.2_{\pm 0.7}$
Exhaustive	$100.00_{\pm 0.00}$	$6.000_{\pm 0.000}$	$12.5_{\pm 3.2}$	$98.50_{\pm 2.20}$	$4.850_{\pm 0.250}$	$10.5_{\pm 3.0}$	$98.20_{\pm 2.10}$	$4.150_{\pm 0.220}$	$7.9_{\pm 2.7}$	$95.80_{\pm 0.91}$	$3.342_{\pm 0.053}$	$6.0_{\pm 0.6}$
RLCP	$95.80_{\pm 3.80}$	$2.910_{\pm 0.520}$	$9.2_{\pm 8.2}$	$95.10_{\pm 3.20}$	$2.850_{\pm 0.480}$	$8.5_{\pm 6.8}$	$94.60_{\pm 2.90}$	$2.850_{\pm 0.460}$	$6.4_{\pm 5.5}$	$93.15_{\pm 2.45}$	$2.540_{\pm 0.305}$	$4.5_{\pm 1.8}$
CluCP	$90.22_{\pm 2.06}$	$3.112_{\pm 0.332}$	$10.7_{\pm 0.9}$	$90.62_{\pm 0.98}$	$2.722_{\pm 0.269}$	$9.9_{\pm 0.7}$	$91.11_{\pm 1.83}$	$2.756_{\pm 0.250}$	$8.9_{\pm 1.1}$	$89.81_{\pm 1.78}$	$2.426_{\pm 0.050}$	$4.9_{\pm 0.8}$
AFCP	$90.67_{\pm 1.44}$	$3.162_{\pm 0.292}$	$6.4_{\pm 0.8}$	$91.30_{\pm 1.24}$	$2.854_{\pm 0.328}$	$7.6_{\pm 0.7}$	$91.62_{\pm 1.71}$	$2.848_{\pm 0.227}$	$7.3_{\pm 0.9}$	$91.77_{\pm 1.23}$	$2.432_{\pm 0.043}$	$4.6_{\pm 0.5}$
FaReG	$92.18_{\pm 1.26}$	$3.536_{\pm 0.388}$	$7.5_{\pm 0.6}$	$92.41_{\pm 1.13}$	$3.178_{\pm 0.288}$	$8.5_{\pm 0.6}$	$92.49_{\pm 1.57}$	$3.141_{\pm 0.238}$	$8.5_{\pm 0.9}$	$92.61_{\pm 0.98}$	$2.579_{\pm 0.029}$	$4.7_{\pm 0.5}$
SGCP	$\mathbf{91.20}_{\pm 1.20}$	$\mathbf{2.050}_{\pm 0.120}$	$\mathbf{6.1}_{\pm 0.9}$	$\mathbf{90.80}_{\pm 1.10}$	$\mathbf{1.820}_{\pm 0.110}$	$\mathbf{5.1}_{\pm 0.8}$	$\mathbf{90.50}_{\pm 1.00}$	$\mathbf{1.710}_{\pm 0.090}$	$\mathbf{4.7}_{\pm 0.8}$	$\mathbf{90.13}_{\pm 0.39}$	$\mathbf{2.225}_{\pm 0.035}$	$\mathbf{4.2}_{\pm 0.7}$

Table A7: Dataset access and preprocessing sources.

Dataset	Source	Access note
Synthetic	Procedurally generated; see Appendix D.1	Generated from the described simulation protocol
Nursery	`https://archive.ics.uci.edu/dataset/76/nursery`	Public UCI dataset
MIMIC-IV	`https://physionet.org/content/mimiciv/3.1/`	Credentialed access via PhysioNet
BACH	`https://zenodo.org/records/3632035`	Public Zenodo record

2048×1536, annotated by two pathologists, and collected from 39 patients, though anonymization prevents complete patient-level tracing. The task is four-class classification with

$$Y \in \{\text{Normal}, \text{Benign}, \text{In Situ Carcinoma}, \text{Invasive Carcinoma}\}.$$

Since demographic sensitive attributes are unavailable, we use the class label as a proxy grouping variable and set $A := Y$. Subgroup reliability is therefore evaluated across disease categories, which naturally differ in visual complexity, intra-class variability, and class-boundary ambiguity. For preprocessing, images are resized to 224×224, augmented during training with random horizontal and vertical flips, and normalized using ImageNet statistics. We train a visual backbone Φ and use the learned representations $X_i = \Phi(I_i) \in \mathbb{R}^d$ for downstream prediction and evaluation.

In this benchmark, heterogeneity is primarily visual rather than demographic. Differences in tissue morphology, staining, lesion structure, and acquisition conditions can induce distinct class-conditional score laws, $F_c(s) := \mathbb{P}(S \le s \mid Y = c)$. Thus, BACH complements the tabular benchmarks by testing whether SGCP can maintain reliable conditional coverage when subgroup structure reflects semantic visual heterogeneity rather than explicit sensitive attributes.

D.2 Implementation Details

Computing resources. All experiments are conducted on a workstation equipped with an Intel i5-14600KF CPU, 32 GB of RAM, and a single NVIDIA GeForce RTX 4090 D GPU. All models are implemented in PyTorch, and each experimental setting is trained with multiple random seeds. The hyperparameters for each dataset are reported in Table A4. Data access and code availability are summarized in Tables A7 and A8, respectively.

Open access to data and code. All external datasets used in our experiments are publicly documented and accessible online, subject to their original access requirements. The synthetic dataset is generated procedurally following Appendix D.1. The code for SGCP is provided anonymously in the supplementary material, and baseline implementations are based on the publicly available repositories listed below.

D.2.1 Baseline implementation details

We compare SGCP with six baselines: Partial Equalized, Exhaustive Equalized, CluCP, AFCP, FaReG, and RLCP. Unless otherwise stated, all methods use the same backbone classifier and softmax-based nonconformity score as SGCP. Hyperparameters are adopted from the original papers or released implementations whenever available.

Table A8: Code availability and baseline implementations.

Method	Code source
SGCP	`https://github.com/MHZhong1225/Restricted-calibration-CP`
Partial Equalized	`https://github.com/yromano/cqr`
RLCP	`https://github.com/rohanhore/RLCP`
CluCP	`https://github.com/tiffanyding/class-conditional-conformal`
AFCP	`https://github.com/FionaZ3696/Adaptively-Fair-Conformal-Prediction`
FaReG	`https://github.com/Xusr1123/FaReG`

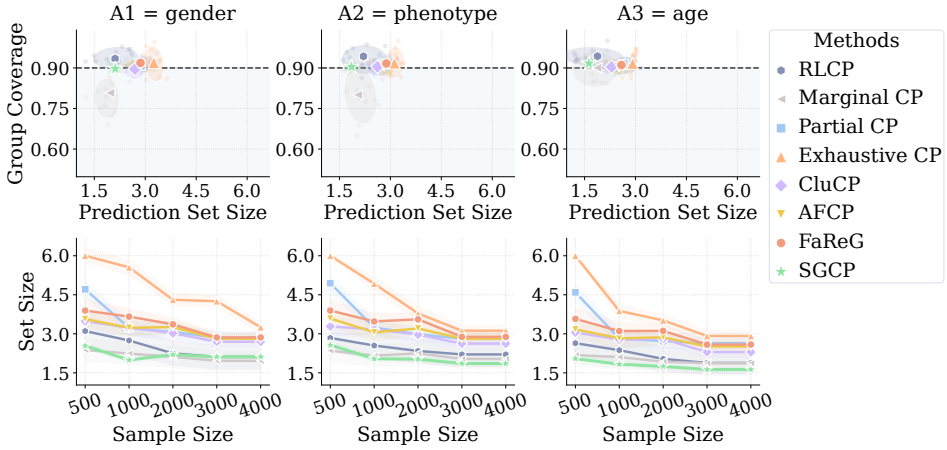


Figure A6: Group coverage on SYN dataset. SGCP remains close to the target coverage across classes while using smaller prediction sets than more conservative baselines.

Partial and Exhaustive Equalized. We extend the equalized-coverage framework of [22] to conformal prediction by applying group-conditional calibration over sensitive groups. Partial Equalized calibrates separately for each value of one sensitive attribute, whereas Exhaustive Equalized calibrates over all intersectional groups induced by the full combination of sensitive attributes. In both cases, the classifier is trained on all samples, and only the calibration step is group-conditional.

CluCP. We implement CluCP following its original clustered conformal pipeline. The clustering fraction γ and the number of clusters M are selected using the heuristic proposed in the paper. Class embeddings are clustered by weighted k-means with the default `KMeans` implementation. Although CluCP targets class-wise rather than subgroup-fair calibration, it provides a useful non-global calibration baseline.

AFCP. We implement AFCP using its original adaptive attribute-selection procedure, directly applying it to the sensitive attributes available in each dataset.

FaReG. We follow the original FaReG representation-learning pipeline: an encoder-decoder first learns latent representations, representation-based low-coverage groups are then identified, and prediction sets are adjusted accordingly. We use the original three-layer MLP encoder and decoder. For Synthetic and Nursery, we adopt the reported hyperparameters: hidden dimensions $[64, 32]$, batch size 500, learning rates 10^{-3} and 10^{-2}, respectively, with $(\beta, \delta, T) = (2.0, 0.3, 20)$ for Synthetic and $(0.1, 0.1, 100)$ for Nursery. For the remaining datasets, we keep the architecture and optimization settings as close as possible to the released implementation.

RLCP. We implement RLCP following its weighted localization procedure. Calibration samples are weighted by their similarity to each test point, and prediction sets are constructed using the corresponding weighted conformal quantile. For all datasets, we use the Gaussian kernel and the bandwidth-selection rule from the original method, without subgroup-specific tuning.

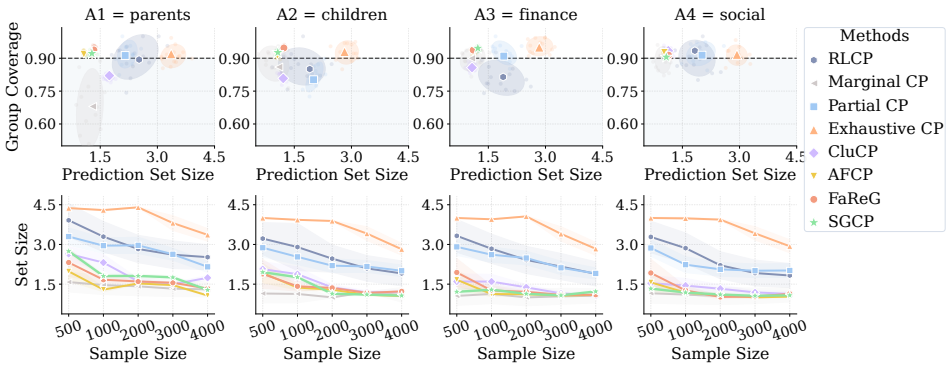


Figure A7: Group coverage on Nursery. Top: subgroup coverage *vs.* average prediction set size at a sample size of 4000, with the dashed line indicating nominal coverage of 0.9. Bottom: prediction set size varies with subgroup sample size. SGCP stays close to the nominal target while remaining among the most efficient methods across subgroup definitions.

Table A9: Performance comparison w.r.t. sample sparsity on the Nursery Dataset (Target $1-\alpha = 0.90$). **Takeaway:** SGCP remains near nominal coverage even in the sparse regime and consistently achieves the smallest subgroup coverage gap, while competing methods improve subgroup reliability only with larger prediction sets or at the cost of under- or over-coverage. And the performance of each specific sensitive group is shown in the Fig. A7.

Method	**N = 500**			**N = 1000**			**N = 2000**			**N = 4000**		
	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓
Marginal	$48.25_{\pm 8.50}$	$1.120_{\pm 0.350}$	$29.5_{\pm 6.5}$	$57.59_{\pm 8.20}$	$1.080_{\pm 0.330}$	$29.1_{\pm 5.2}$	$59.72_{\pm 7.80}$	$1.050_{\pm 0.310}$	$26.2_{\pm 5.8}$	$84.15_{\pm 5.20}$	$2.170_{\pm 0.260}$	$20.4_{\pm 5.2}$
Partial	$84.61_{\pm 4.80}$	$2.850_{\pm 0.310}$	$17.7_{\pm 3.8}$	$86.07_{\pm 4.50}$	$2.620_{\pm 0.280}$	$16.8_{\pm 3.5}$	$87.86_{\pm 4.20}$	$2.450_{\pm 0.250}$	$15.2_{\pm 3.2}$	$88.97_{\pm 3.80}$	$1.850_{\pm 0.200}$	$12.1_{\pm 2.5}$
Exhaustive	$100.00_{\pm 0.00}$	$4.000_{\pm 0.000}$	$19.5_{\pm 0.1}$	$98.00_{\pm 0.10}$	$3.980_{\pm 0.001}$	$18.8_{\pm 0.2}$	$98.00_{\pm 0.12}$	$3.979_{\pm 0.001}$	$17.5_{\pm 0.8}$	$96.80_{\pm 3.10}$	$2.850_{\pm 0.220}$	$15.5_{\pm 3.2}$
RLCP	$82.45_{\pm 14.20}$	$2.250_{\pm 0.580}$	$18.5_{\pm 9.5}$	$83.36_{\pm 13.80}$	$1.720_{\pm 0.550}$	$17.8_{\pm 6.2}$	$83.31_{\pm 11.50}$	$1.480_{\pm 0.510}$	$16.5_{\pm 5.8}$	$88.67_{\pm 6.20}$	$1.220_{\pm 0.410}$	$13.5_{\pm 3.8}$
CluCP	$86.26_{\pm 5.61}$	$1.506_{\pm 0.265}$	$17.3_{\pm 4.5}$	$83.80_{\pm 3.59}$	$1.516_{\pm 0.265}$	$17.1_{\pm 2.9}$	$84.42_{\pm 2.47}$	$1.261_{\pm 0.096}$	$16.9_{\pm 1.8}$	$85.08_{\pm 1.60}$	$1.121_{\pm 0.064}$	$12.8_{\pm 1.3}$
AFCP	$92.78_{\pm 4.81}$	$1.551_{\pm 0.345}$	$18.7_{\pm 3.9}$	$94.85_{\pm 3.69}$	$\mathbf{1.163}_{\pm 0.074}$	$16.4_{\pm 2.8}$	$93.12_{\pm 1.33}$	$\mathbf{1.092}_{\pm 0.031}$	$15.6_{\pm 1.1}$	$92.85_{\pm 0.84}$	$\mathbf{1.085}_{\pm 0.024}$	$14.1_{\pm 0.6}$
FaReG	$93.77_{\pm 3.66}$	$1.879_{\pm 0.515}$	$17.4_{\pm 2.9}$	$92.47_{\pm 3.04}$	$1.346_{\pm 0.146}$	$16.2_{\pm 2.3}$	$94.37_{\pm 1.44}$	$1.257_{\pm 0.027}$	$15.4_{\pm 1.1}$	$93.21_{\pm 0.62}$	$1.230_{\pm 0.012}$	$13.3_{\pm 0.6}$
SGCP	$\mathbf{92.72}_{\pm 1.50}$	$\mathbf{1.320}_{\pm 0.080}$	$\mathbf{15.1}_{\pm 1.5}$	$\mathbf{92.10}_{\pm 1.20}$	$1.210_{\pm 0.060}$	$\mathbf{14.6}_{\pm 1.2}$	$\mathbf{91.97}_{\pm 1.00}$	$1.120_{\pm 0.050}$	$\mathbf{12.3}_{\pm 1.0}$	$\mathbf{91.65}_{\pm 0.70}$	$1.089_{\pm 0.035}$	$\mathbf{11.8}_{\pm 0.7}$

E Additional Experiments

E.1 Ablation Study

Table A5 studies the contribution of each component in SGCP. The default model achieves nominal marginal coverage on all datasets and obtains the best or tied-best CovGap on three out of four datasets, while maintaining the smallest or near-smallest prediction sets. This suggests that the full objective provides the most favorable balance between validity, subgroup reliability, and efficiency.

Removing posterior stochasticity consistently degrades subgroup reliability. The deterministic variant increases CovGap on all datasets, indicating that posterior averaging is useful for stabilizing the local score law rather than committing each sample to a single calibration state. Replacing the learned calibration state with feature-space grouping also leads to larger CovGap and larger prediction sets, especially on MIMIC and BACH. This supports our motivation that similarity in raw feature space is not necessarily the right notion of locality for conformal calibration; what matters is similarity in score behavior.

The loss ablations further clarify the role of each training signal. Without the rank-uniformization loss, coverage becomes less stable and CovGap increases substantially, showing that aligning transformed scores to a common percentile scale is central to subgroup reliability. Removing the density-fitting term also worsens the tradeoff between coverage and efficiency, often producing larger prediction sets, which suggests that likelihood-based fitting helps the learned local law remain faithful to the observed score distribution. Finally, removing the mutual-information term leads to less reliable subgroup calibration and higher variance across runs, consistent with component underuse or less informative

Table A10: Performance comparison w.r.t. sample sparsity on MIMIC-IV (target $1 - \alpha = 0.90$). Each method is run 20 times. **Takeaway:** On the MIMIC-IV dataset, SGCP consistently achieves near-nominal coverage while maintaining the lowest subpopulation disparity (CovGap) in sparse settings ($N = 2000, 5000$). Even with the full dataset, SGCP provides a highly efficient prediction set with excellent fairness metrics compared to conventional baselines.

Method	**N = 2000**			**N = 5000**			**Full (All Samples)**		
	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓
Marginal	$90.06_{\pm 10.14}$	$1.074_{\pm 0.211}$	$34.4_{\pm 10.9}$	$90.79_{\pm 10.62}$	$\mathbf{1.026}_{\pm 0.087}$	$29.6_{\pm 9.3}$	$90.79_{\pm 10.62}$	$1.226_{\pm 0.087}$	$27.1_{\pm 7.1}$
Partial	$95.73_{\pm 3.62}$	$1.564_{\pm 0.243}$	$11.0_{\pm 5.6}$	$94.50_{\pm 4.54}$	$1.335_{\pm 0.207}$	$9.2_{\pm 4.0}$	$94.50_{\pm 4.54}$	$1.258_{\pm 0.140}$	$8.0_{\pm 3.0}$
Exhaustive	$100.00_{\pm 0.00}$	$3.000_{\pm 0.000}$	$18.7_{\pm 8.0}$	$98.86_{\pm 5.21}$	$2.989_{\pm 0.004}$	$21.4_{\pm 5.3}$	$96.45_{\pm 4.22}$	$2.800_{\pm 0.142}$	$19.7_{\pm 3.2}$
RLCP	$95.10_{\pm 2.40}$	$1.650_{\pm 0.120}$	$4.2_{\pm 2.6}$	$93.20_{\pm 2.10}$	$1.480_{\pm 0.150}$	$3.5_{\pm 1.4}$	$92.50_{\pm 2.80}$	$1.350_{\pm 0.100}$	$1.7_{\pm 0.8}$
CluCP	$95.11_{\pm 1.92}$	$1.654_{\pm 0.192}$	$4.5_{\pm 1.5}$	$93.19_{\pm 1.28}$	$1.422_{\pm 0.059}$	$3.5_{\pm 1.0}$	$92.22_{\pm 1.28}$	$1.302_{\pm 0.059}$	$3.5_{\pm 0.8}$
AFCP	$91.81_{\pm 2.51}$	$1.400_{\pm 0.075}$	$3.9_{\pm 1.0}$	$91.40_{\pm 1.12}$	$1.114_{\pm 0.045}$	$3.3_{\pm 0.8}$	-	-	-
FaReG	$95.78_{\pm 2.09}$	$1.601_{\pm 0.078}$	$4.1_{\pm 1.2}$	$92.58_{\pm 1.05}$	$1.357_{\pm 0.050}$	$3.0_{\pm 1.1}$	$90.23_{\pm 0.09}$	$1.174_{\pm 0.001}$	$\mathbf{0.5}_{\pm 0.0}$
SGCP	$\mathbf{90.12}_{\pm 1.91}$	$\mathbf{1.060}_{\pm 0.068}$	$\mathbf{3.3}_{\pm 0.7}$	$\mathbf{90.48}_{\pm 1.36}$	$1.116_{\pm 0.050}$	$\mathbf{2.6}_{\pm 0.5}$	$\mathbf{90.17}_{\pm 0.07}$	$\mathbf{1.107}_{\pm 0.001}$	$0.8_{\pm 0.0}$

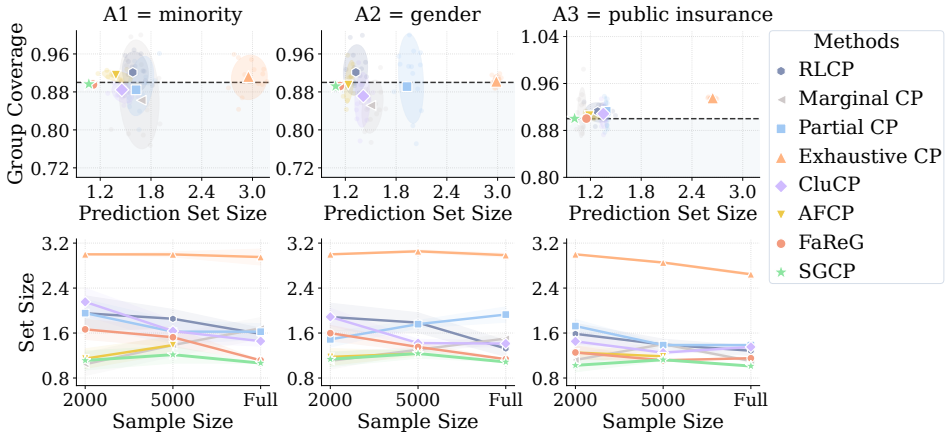


Figure A8: Group coverage on MIMIC-IVdataset. SGCP remains close to the target coverage across classes while using smaller prediction sets than more conservative baselines.

stochastic grouping. Overall, the ablation results support the design of SGCP: subgroup reliability improves when local score laws are learned through both distributional fitting and rank-based normalization, and when calibration evidence is shared through informative stochastic components.

E.2 Additional Biomedical Benchmarks

To further assess the generalizability of SGCP across biomedical imaging tasks, we conduct additional experiments on two datasets, HAM10000 and MIMIC-CXR. These experiments extend our evaluation to dermatology and chest radiography, respectively.

Datasets. HAM10000 [25] is a public dermoscopic image dataset containing 10,015 images from seven diagnostic categories. The associated metadata include sex, age, and anatomical lesion site, enabling subgroup evaluation based on individual attributes and their intersections. MIMIC-CXR [14, 15] is a large-scale clinical chest radiograph dataset with associated demographic and encounter-level information. We use it to evaluate thoracic disease prediction and intersectional subgroup reliability based on available patient-level attributes.

Results. At 90% nominal coverage, Table A12shows that SGCP achieves 90.30% coverage and the lowest intersectional CovGap of 5.99 on HAM10000. Its average set size is 1.36, comparable to Marginal CP and Partial, and smaller than AFCP and FaReG. On MIMIC-CXR (as shown in Table A13), SGCP achieves 90.25% coverage with an average set size of 1.39, matching Marginal CP. Its intersectional CovGap of 1.21 is close to those of Marginal CP and AFCP, at 1.24 and 1.14,

Table A11: Performance comparison w.r.t. sample sparsity on BACH (target $1-\alpha = 0.90$). Each method is run 20 times. **Takeaway:** Under extreme sparsity, most methods become conservative, but with the increase in sample size, SGCP recovers the fastest, with near-nominal coverage, a smaller prediction set, and lower subpopulation disparity, outperforming other baseline methods. The performance of each individual subgroup is further shown in Fig. 4.

Method	N = 50			N = 150			N = 200			N = 300		
	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓	Cov. ↑	AvgSize ↓	CovGap ↓
Marginal	$89.34_{\pm 3.35}$	$3.700_{\pm 0.474}$	$19.0_{\pm 2.1}$	$90.50_{\pm 3.86}$	$3.604_{\pm 0.494}$	$17.6_{\pm 2.0}$	$90.69_{\pm 15.75}$	$3.545_{\pm 0.445}$	$15.9_{\pm 2.9}$	$89.20_{\pm 9.64}$	$2.894_{\pm 0.267}$	$15.7_{\pm 1.7}$
Partial	$98.33_{\pm 2.66}$	$3.735_{\pm 0.316}$	$7.8_{\pm 1.0}$	$97.45_{\pm 3.25}$	$3.711_{\pm 0.290}$	$\mathbf{6.2}_{\pm 0.6}$	$97.17_{\pm 11.56}$	$3.460_{\pm 0.258}$	$7.5_{\pm 0.7}$	$96.57_{\pm 8.08}$	$3.385_{\pm 0.299}$	$7.6_{\pm 1.6}$
Exhaustive	$100.00_{\pm 0.00}$	$6.000_{\pm 0.000}$	$10.0_{\pm 0.0}$	$98.50_{\pm 2.20}$	$4.850_{\pm 0.250}$	$10.5_{\pm 3.0}$	$98.20_{\pm 2.10}$	$4.150_{\pm 0.220}$	$7.9_{\pm 2.7}$	$95.85_{\pm 0.91}$	$3.342_{\pm 0.053}$	$6.0_{\pm 0.6}$
RLCP	$91.74_{\pm 3.06}$	$\mathbf{2.855}_{\pm 0.440}$	$8.2_{\pm 3.6}$	$92.25_{\pm 3.80}$	$\mathbf{2.763}_{\pm 0.422}$	$7.9_{\pm 3.8}$	$91.99_{\pm 10.88}$	$2.611_{\pm 0.411}$	$6.4_{\pm 3.8}$	$91.88_{\pm 10.56}$	$2.311_{\pm 0.319}$	$8.1_{\pm 2.5}$
CluCP	$97.80_{\pm 3.14}$	$3.936_{\pm 0.199}$	$8.7_{\pm 1.5}$	$96.60_{\pm 3.29}$	$3.882_{\pm 0.180}$	$8.1_{\pm 1.9}$	$98.34_{\pm 0.50}$	$3.985_{\pm 0.110}$	$10.0_{\pm 2.6}$	$95.91_{\pm 3.59}$	$3.257_{\pm 0.220}$	$10.7_{\pm 2.6}$
AFCP	$100.00_{\pm 0.50}$	$4.000_{\pm 0.120}$	$10.0_{\pm 1.2}$	$96.50_{\pm 2.64}$	$3.817_{\pm 0.116}$	$8.3_{\pm 1.3}$	$95.80_{\pm 2.39}$	$3.447_{\pm 0.297}$	$7.5_{\pm 1.2}$	$95.00_{\pm 2.60}$	$2.830_{\pm 0.237}$	$6.3_{\pm 1.7}$
FaReG	$99.00_{\pm 2.00}$	$3.946_{\pm 0.153}$	$9.3_{\pm 1.3}$	$98.40_{\pm 1.78}$	$3.914_{\pm 0.212}$	$8.4_{\pm 1.8}$	$96.80_{\pm 1.93}$	$3.769_{\pm 0.250}$	$7.7_{\pm 1.5}$	$93.60_{\pm 3.40}$	$2.959_{\pm 0.239}$	$7.4_{\pm 1.5}$
SGCP	$96.44_{\pm 2.61}$	$3.842_{\pm 0.136}$	$\mathbf{6.8}_{\pm 2.2}$	$92.63_{\pm 4.04}$	$3.344_{\pm 0.240}$	$\mathbf{5.0}_{\pm 1.58}$	$92.22_{\pm 2.39}$	$\mathbf{2.370}_{\pm 0.197}$	$\mathbf{4.6}_{\pm 2.1}$	$\mathbf{92.00}_{\pm 2.70}$	$\mathbf{2.224}_{\pm 0.215}$	$\mathbf{6.2}_{\pm 1.1}$

Table A12: Performance comparison on the additional biomedical dataset HAM10000 at 90% nominal coverage. **Takeaway:** SGCP achieves the lowest intersectional CovGap while maintaining near-nominal coverage and compact prediction sets.

Method	Coverage (%)	AvgSize	Intersection CovGap
Marginal CP	90.25 ± 0.88	1.35 ± 0.05	7.30 ± 0.62
Partial	90.24 ± 0.90	1.34 ± 0.05	7.40 ± 0.54
AFCP	90.44 ± 0.86	1.64 ± 0.10	6.20 ± 0.49
FaReG	93.02 ± 1.09	1.93 ± 0.15	8.00 ± 1.09
SGCP	90.30 ± 0.88	1.36 ± 0.05	$\mathbf{5.99 \pm 0.84}$

respectively, and substantially lower than FaReG at 3.78. AFCP achieves a slightly lower CovGap, but with larger prediction sets.

Together, SGCP achieves the lowest intersectional CovGap on HAM10000 and competitive subgroup reliability on MIMIC-CXR, while maintaining compact prediction sets on both datasets. These results extend our evaluation to dermatology and chest radiography, suggesting that the findings on the original benchmarks are not limited to a single clinical task or imaging modality.

E.3 Sensitivity to Hyperparameters and Subgroup Configurations

E.3.1 Effect of sample sparsity

Tables A6–A11 study how performance changes as the effective sample size varies. As the effective sample size decreases, methods based on explicit subgroup equalization or localization tend to become increasingly conservative, while global calibration remains efficient but suffers from large subgroup disparity. By contrast, SGCP degrades more gracefully in the sparse regime and benefits more consistently from additional data as sample size grows. This pattern is strongest on the synthetic benchmark and on BACH, where subgroup heterogeneity is more structured, and remains visible on MIMIC-IV, where SGCP is particularly competitive in the sparse and moderate sample regimes.

For a clearer comparison, we also provide separate box plots for each metric across methods and sample sizes. Figures A9, A10, and 5 show the distributions of overall coverage, average prediction set size, and subgroup coverage gap under varying sample sparsity. Each box summarizes 20 independent trials with different random seeds for a given dataset and sample-size setting. Across datasets, SGCP remains near the nominal target with competitive set sizes and consistently low subgroup disparity, especially in sparse regimes where competing methods become more conservative or more variable.

E.3.2 Effect of the number of latent components

Figure A11 visualizes the learned component densities on SYN under different choices of K. When K is small, the model learns a few distinct score law patterns that separate low score and high score calibration regimes. As K increases, additional components tend to overlap with existing ones or

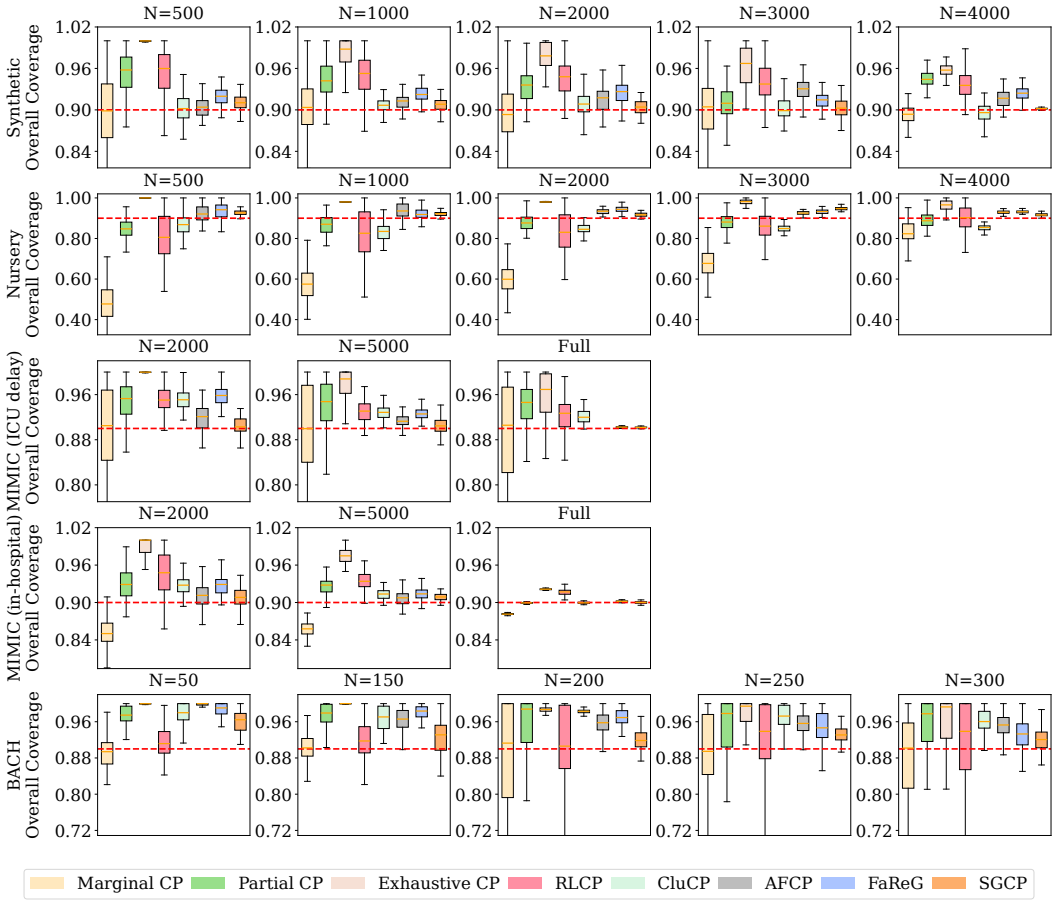


Figure A9: Distribution of overall coverage over repeated runs under varying sample sparsity. The dashed line marks the nominal target coverage of 0.9. **Takeaway:** SGCP remains consistently close to the target across datasets and sample sizes, whereas global calibration may under-cover in sparse regimes and stronger equalization baselines often become overly conservative.

Table A13: Performance comparison on the additional biomedical dataset MIMIC-CXR. **Takeaway:** SGCP matches Marginal CP in average set size and achieves competitive intersectional reliability.

Method	Coverage (%)	AvgSize	Intersection CovGap
Marginal CP	90.27 ± 0.70	1.39 ± 0.34	1.24 ± 0.34
Partial	90.49 ± 0.63	1.41 ± 0.02	1.35 ± 0.33
AFCP	90.31 ± 0.86	1.49 ± 0.02	$\mathbf{1.14 \pm 0.32}$
FaReG	93.78 ± 0.59	1.62 ± 0.03	3.78 ± 0.59
SGCP	90.25 ± 0.72	1.39 ± 0.02	1.21 ± 0.33

become nearly flat, suggesting that the dominant calibration structure has already been captured by a smaller number of components. This behavior indicates that K controls the granularity of stochastic sharing rather than simply increasing model capacity. In practice, overly large K may introduce redundant components, while a moderate K provides a compact set of reusable score law building blocks for constructing sample dependent local score laws.

Figure A12 shows a similar sensitivity pattern on MIMIC-IV, but with a different score structure. Most learned components concentrate near small nonconformity scores, reflecting the strong mass of the empirical score distribution in the low score region. Increasing K refines this region and introduces broader components that capture the remaining tail behavior. However, when K becomes

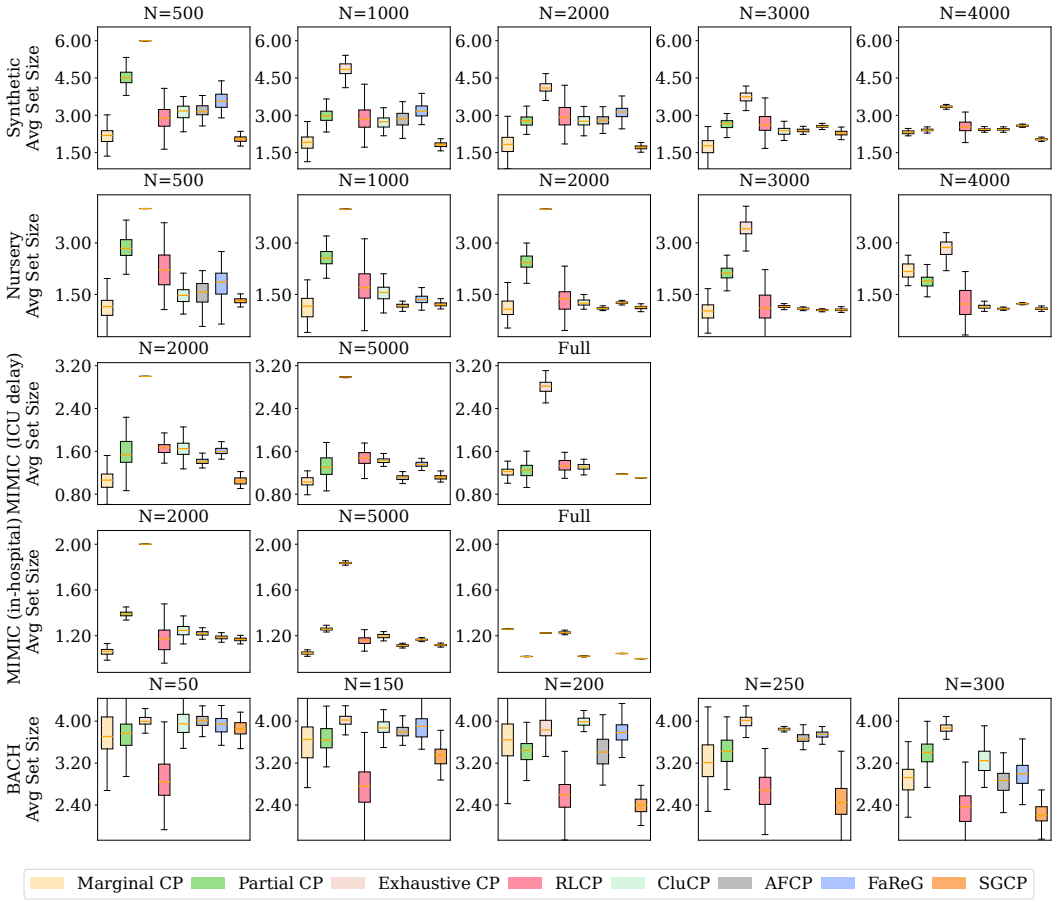
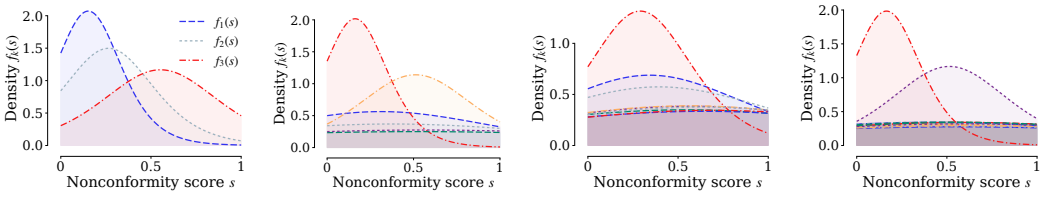


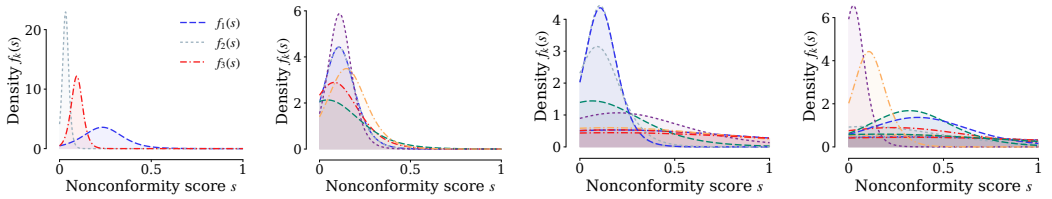
Figure A10: Distribution of average prediction set size over repeated runs under varying sample sparsity. Smaller values indicate better efficiency. **Takeaway:** SGCP remains among the most efficient methods across datasets and sample sizes, often matching or improving upon the smallest prediction sets while avoiding the severe conservatism of explicit equalization and localized baselines.



(a) K=3 (b) K=6 (c) K=9 (d) K=12

Figure A11: Effect of K on learned latent component densities for SYN. Increasing K refines the score law components, but large K introduces overlapping or redundant densities.

large, several components overlap substantially or become weakly differentiated. This suggests that additional components mainly refine existing calibration patterns rather than uncover entirely new score regimes. Together, the SYN and MIMIC-IV results show that a moderate number of latent components is sufficient to capture the dominant score law structure, while larger K may lead to redundant calibration sharing.



(a) K=3 (b) K=6 (c) K=9 (d) K=12

Figure A12: Effect of K on learned latent component densities for MIMIC-IV. Larger K refines the dominant low score region and tail behavior, but also introduces increasingly overlapping components.

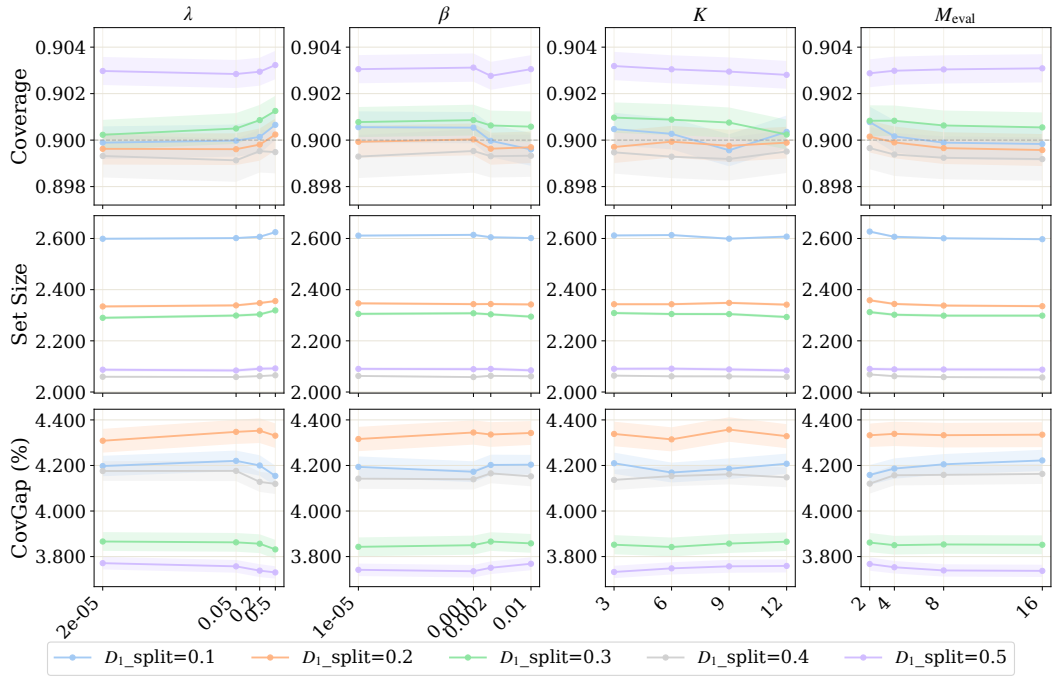


Figure A13: Hyperparameter sensitivity of SGCP on SYN. We vary λ_{nll}, λ_{bal}, K, and M_{eval}, and report coverage, average set size, and subgroup coverage gap under different D_1 split ratios. Shaded bands show one standard deviation. **Takeaway:** SGCP is largely insensitive to these hyperparameters, while larger samples for score-law learning reduce covgap.

E.3.3 Hyperparameter sensitivity

Figures A13 and A14 examine the sensitivity of SGCP w.r.t. several key hyperparameters, including λ, β, the number of latent calibration components K, and the number of evaluation latent samples M_{eval}, under different D_1 split ratios. Across both datasets, the hyperparameter curves are generally flat, indicating that the performance of SGCP does not rely on delicate tuning.

On the SYN, this stability is particularly clear: coverage remains close to the nominal target across all tested settings, while the changes in average prediction set size and subgroup coverage gap are modest. The most visible effect comes from the D_1 split ratio, where allocating more data to score law learning consistently improves subgroup reliability and reduces prediction set size. On MIMIC-IV, it show a similar pattern. Coverage remains tightly concentrated around the nominal level across all tested settings, with modest changes in AvgSize and CovGap. Compared with hyperparameter choices, allocation training data to local score law learning has a more visible effect on the final tradeoff. Overall, these results suggest that the gains of SGCP stem primarily from its structured local score law design rather than from aggressive hyperparameter tuning.

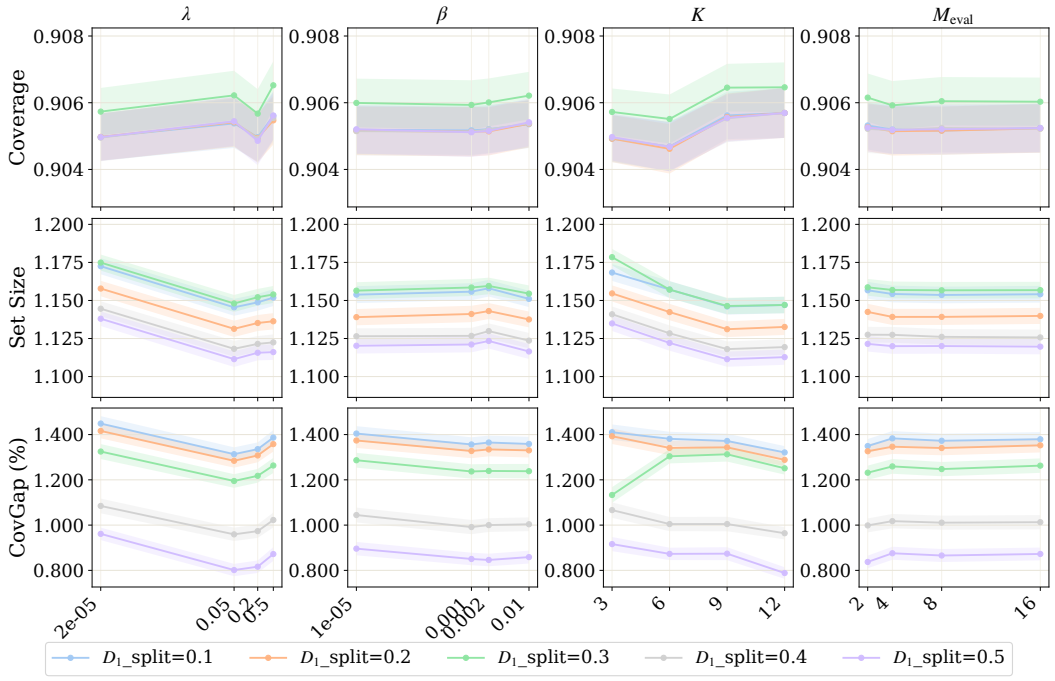


Figure A14: Hyperparameter sensitivity of SGCP on MIMIC-IV (ICU delay). **Takeaway:** SGCP is largely insensitive to these hyperparameters. Coverage remains close to the nominal level, and most of the variation in efficiency and subgroup reliability is driven by the D_1 split ratio rather.

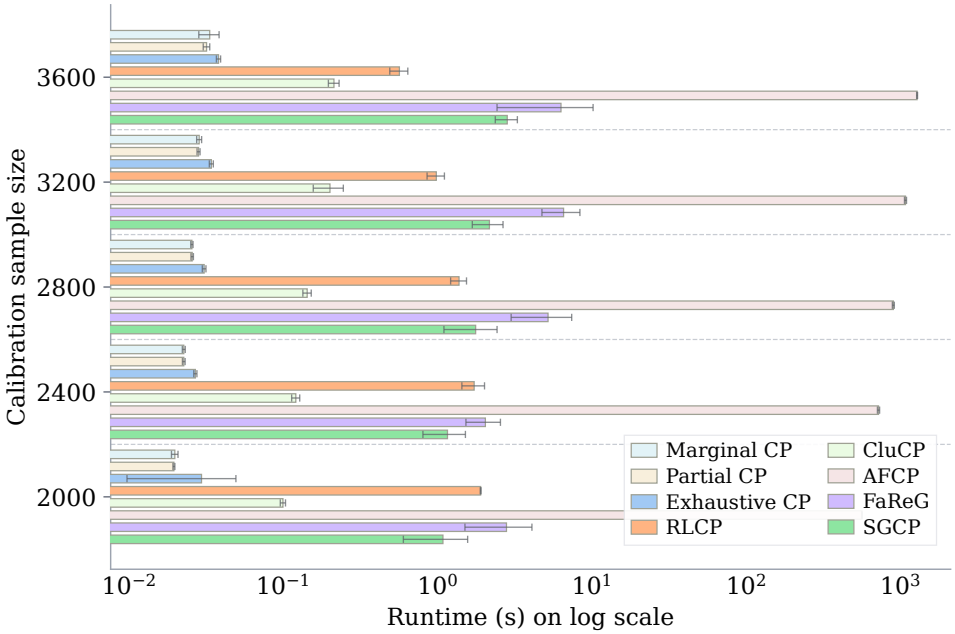


Figure A15: Wall clock time comparison w.r. calibration sample. We report the wall-clock time for each method, averaged over repeated runs; error bars indicate one standard deviation. The horizontal axis is shown on a log scale. **Takeaway:** SGCP has moderate computational cost and is faster than AFCP, while remaining competitive with other subgroup aware baselines across all sample sizes.

E.4 Wall-Clock Time

We compare the wall-clock time of the full SGCP pipeline with all baselines under varying calibration sample sizes. Figure A15 reports the average runtime over repeated runs, with error bars indicating one standard deviation. Marginal CP and Partial CP are the fastest methods, as they require only global calibration. RLCP and SGCP incur additional cost from sample-dependent calibration, but remain moderate and scale smoothly with sample size. AFCP is the most expensive baseline by a large margin, reflecting its repeated adaptive subgroup selection and conformal recalibration, while CluCP and FaReG lie in between. Overall, SGCP provides a favorable computational tradeoff, remaining practical across all tested sample sizes while achieving stronger subgroup reliability than simpler global or rigid group based baselines.

E.5 Limitations

Our study has several limitations. First, the conformal coverage guarantee relies on exchangeability between calibration and test data, which may be violated under temporal, institutional, or other distribution shifts. Establishing coverage for a new cohort therefore requires recalibration using target data that satisfy this assumption. Nevertheless, because SGCP separates score-law learning from conformal calibration, such adaptation does not necessarily require retraining the entire model. As demonstrated in our temporal transfer experiment, the source-trained backbone, router, and component laws can remain frozen, while only the final conformal cutoff is recalibrated on target data. Second, SGCP's subgroup coverage guarantee is mismatch-dependent rather than exact distribution-free conditional coverage. Evaluation on recorded sensitive attributes cannot rule out coverage failures in unobserved or extremely sparse subgroups. Moreover, the learned latent components capture recurring calibration regimes and should not be interpreted as distinct patient populations. SGCP also inherits limitations of the predictive backbone and nonconformity score. Finally, our experiments are retrospective and focus on multiclass classification. Prospective validation in clinical workflows, extensions to other prediction settings such as regression, and evaluation under real-world deployment constraints remain important directions for future work.

F Supplement to Related Work

Group-conditional and fairness-aware conformal prediction. A substantial body of work studies fairness-aware conformal prediction through coverage parity across groups that are either specified in advance or adaptively selected from data. Vadlamani et.al [26] develop a unified post-hoc framework that extends classical fairness notions to prediction sets by adjusting conformity thresholds in a fairness-aware manner. AFCP [32] study conformal classification with equalized coverage for adaptively selected groups, selecting features that reflect potential model limitations or biases in order to balance predictive efficiency and group reliability. FAREG [30] goes beyond predefined sensitive groups by learning representation-based subgroups and adaptively promoting equalized coverage for populations that would otherwise receive unfair treatment. SAGCCI [11] examines the tradeoff between fairness and efficiency by clustering protected groups with similar conformal score distributions and by using surrogate outcomes to improve threshold estimation. This is particularly close in spirit to our work in its use of score-distribution similarity to share calibration information. The key distinction lies in the granularity of the sharing mechanism. SAGCCI shares calibration information across clusters of observed protected groups, whereas SGCP learns a sample-dependent stochastic mixture over latent calibration components and thereby induces a local score law for each individual sample. Consequently, the calibration-sharing structure learned by SGCP need not align with protected-group labels and does not rely on surrogate outcomes.

Structured sharing and learned calibration groups. A related line of work improves upon a single global calibration rule by sharing calibration evidence across structured partitions. Clustered conformal prediction [8] groups labels with similar score distributions in many-class settings and performs calibration at the cluster level, thereby interpolating between standard conformal prediction and fully class-conditional conformal prediction. Conceptually similar ideas also arise beyond conformal prediction. For instance, VIR [29] smooths probabilistic uncertainty estimates across neighboring label regions in imbalanced regression. Together, these works underscore the value of structured sharing: calibration and uncertainty estimation can often benefit from borrowing information across statistically related regions, rather than treating all samples globally or calibrating each group in

isolation. Another line of work enlarges or learns the group structure used for conditional calibration. Kandinsky Conformal Prediction [3] generalizes group-conditional conformal inference to overlapping and fractional group functions defined on (X, Y), thereby expanding the collection of groups over which weighted coverage deviations can be controlled. Recent work on identifying homogeneous and interpretable groups for conformal prediction likewise uses nonconformity-score behavior to learn explicit partitions of the input space before applying group-conditional calibration [20]. These methods are closely related to ours in their shared emphasis on calibration-relevant structure beyond predefined sensitive attributes.

Our work adopts a complementary perspective. Instead of specifying a family of group functions or assigning each sample to a hard discovered partition, SGCP learns a sample-dependent local score law through stochastic mixing over shared latent calibration components. This enables calibration information to be shared according to learned score behavior, while the final prediction set is still obtained through standard split conformal calibration applied to transformed scores.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are supported by the proposed method, theoretical results, and empirical evaluation. The scope of the claims is limited to subgroup reliability for conformal prediction under the benchmark settings studied in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix E.5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper includes formal propositions with stated assumptions and provides the corresponding proofs in the Appendix A.1. The theoretical results are used to support the proposed calibration procedure.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides the information needed to reproduce the main experiments, including dataset construction, preprocessing, train/calibration/test splits, model architecture, baseline implementations, and hyperparameters (see Appendix B). Code is also provided to support reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides code with instructions for reproducing the main experimental results. Public datasets are cited and described, while restricted-access data such as MIMIC-IV are used under their original access requirements.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental protocol, including dataset construction, data splits, sensitive attributes used for evaluation, backbone models, baseline implementations, optimizer settings, and dataset-specific hyperparameters. Additional implementation details are provided in the Appendix D.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main results are reported over multiple runs with error bars. The paper reports mean performance and variability across runs for coverage, average prediction set size, and subgroup coverage gap.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the computational resources and running time required for the experiments (see Appendix D.2 and E.4). These details provide practical guidance for reproducing the empirical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses synthetic, public, or credentialed de-identified datasets and follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the potential positive impact of improving subgroup reliability in uncertainty quantification, especially in high-stakes settings such as healthcare.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk generative models, scraped datasets, or other assets with substantial dual-use misuse potential. The clinical data used in the experiments are de-identified and accessed under the original dataset governance procedures.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources for all datasets and baseline methods, including Nursery, MIMIC-IV, BACH, and prior conformal prediction baselines. Dataset access terms and licenses are respected according to their original providers.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper releases code to support reproduction of the proposed method and experiments, together with documentation and instructions. The synthetic benchmark is generated procedurally and described in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not conduct new human-subject experiments. The clinical benchmark uses de-identified MIMIC-IV data under its established access and data-use requirements.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.