Missingness-Aware Conformal Prediction Under Cross-Hospital Distribution Shift

Anonymous Authors[1]

Abstract

Split conformal prediction (CP) guarantees marginal coverage under exchangeability, but cross-hospital deployment breaks this assumption because missingness patterns differ across hospitals. We show that pooled calibration can hide clinically important subgroup coverage gaps: on GOSSIS, standard CP attains 90% marginal coverage yet covers a low-missingness subgroup at only 82%. We decompose the subgroup gap into a calibration heterogeneity term η_k and a within-group shift term δ_k, showing that the Mondrian bound removes η_k and is tighter when η_k exceeds the finite-sample grouping cost. We then introduce a label-free selection rule that chooses the missingness variable with the largest cross-site missingness shift and calibrates within its two subgroups. On GOSSIS, our method halves the maximum subgroup gap from 0.028–0.044 to 0.015–0.021 with less than 1% change in set size; on MIMIC-IV, it reduces the gap from 0.049–0.075 to 0.026–0.046. Subgroup assignment is invariant to model updates and remains stable under deployment-time stress tests where optimization-based baselines degrade substantially.

1. Introduction

Clinical risk models are increasingly deployed across hospitals, but their reliability degrades when deployment sites follow different measurement practices (Pollard et al., 2018). Conformal prediction (CP) converts point predictions into prediction sets with nominal coverage $1-\alpha$ (Vovk et al., 2005), but split CP relies on exchangeability, which cross-hospital deployment violates through distribution shift in missingness patterns.

In ICU data, missingness is clinically meaningful: whether a laboratory test is observed depends on acuity, workflow, and local measurement policy (Sharafoddini et al., 2019). Prior work shows that even under exchangeability, different missingness masks can induce different conformity score distributions (Zaffran et al., 2023).We observe this on hospital-disjoint ICU mortality prediction: standard CP achieves 90% marginal coverage on GOSSIS/eICU while covering a low-missingness subgroup at only 82% and over-covering a high-missingness subgroup at 94%.

Our key insight is that missingness is useful twice: first as a diagnostic of score heterogeneity, and then as a portable grouping variable for Mondrian calibration. We formalize this with a finite-sample decomposition of subgroup coverage error into a pooled-versus-group mismatch term η_k and a residual within-group shift term δ_k . This decomposition explains when groupwise calibration should help, and it also clarifies what it cannot fix: Mondrian conditioning removes heterogeneity from pooled calibration, but not residual within-group shift.

Building on this view, we propose a label-free, model-agnostic selection rule that chooses the missingness indicator with the largest cross-site distributional shift and calibrates within its observed/missing groups. The grouping depends only on the missingness mask, not on predicted risk, labels, or held-out test data. This makes the resulting calibration protocol portable across model updates, which is operationally important in clinical deployment.

Contributions. (1) We diagnose subgroup coverage gaps in standard CP under hospital-disjoint shift, driven by missingness heterogeneity. (2) We prove a finite-sample bound in which the η_k term is absent under Mondrian calibration, yielding a tighter subgroup guarantee whenever η_k exceeds the grouping cost $1/(n_k{+}1) - 1/(n{+}1)$. (3) We introduce a label-free variable selection rule whose subgroup assignment is model-agnostic and invariant to model updates. (4) We validate on simulations, GOSSIS, and MIMIC-IV with stress and model-swap tests.

2. Related Work

Split CP provides marginal coverage under exchangeability (Papadopoulos et al., 2002; Vovk et al., 2005), while

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

exact distribution-free conditional coverage is impossible without trivially large prediction sets (Barber et al., 2021). Practical relaxations include Mondrian CP, feature-based approximations (Gibbs et al., 2025), and learned partitions (Martínez Gil et al., 2024). Our focus is different: we use missingness-defined grouping to improve subgroup reliability under cross-hospital deployment shift.

Under shift, weighted and non-exchangeable CP methods address marginal coverage degradation (Tibshirani et al., 2019; Barber et al., 2023; Liu et al., 2024). For missing data, Zaffran et al. (2023) and Fan et al. (2024) study mask-conditional validity within a single calibration distribution. Clinical ML work shows that missingness is informative and varies across hospitals (Che et al., 2018; Sharafoddini et al., 2019; Tipirneni & Reddy, 2021; Rockenschaub et al., 2024). We combine these strands by using missingness as a deployment descriptor for subgroup calibration under cross-site shift.

3. Background and Theory

3.1. Setup

Consider samples $\{(X_i, M_i, Y_i)\}_{i=1}^N$ with $X_i \in \mathbb{R}^p$, $Y_i \in \{0,1\}$, and missingness mask $M_i \in \{0,1\}^d$ ($M_{i,j} = 1$ if feature j is unobserved). Let $\hat{f}$ be a fixed base model and Φ a fixed imputation map; since Φ does not change across our analyses, we suppress it and write $\hat{f}(X_i, M_i)$ for brevity. The conformity score $S_i = s(\hat{f}(X_i, M_i), Y_i)$ can be any real-valued function satisfying larger-is-worse; our theory applies to general nonconformity scores, while all experiments use the classification residual $s_i = 1 - \hat{p}_i[y_i]$. Data are split by hospital into disjoint training, selection, calibration (n samples), and test sets.

Standard split CP. The global threshold $\hat{q}$ is the $\lceil (1-\alpha)(n+1) \rceil$-th order statistic of calibration scores; the prediction set is $C_{\text{std}}(x) = \{y : s(\hat{f}(x), y) \le \hat{q}\}$.

Mondrian CP. A grouping function $G : \{0,1\}^d \to [K]$ partitions calibration samples by missingness mask. Per-group thresholds $\hat{q}_k$ are computed within each group; a test point with mask M uses $C_{\text{mon}}(x, M) = \{y : s(\hat{f}(x), y) \le \hat{q}_{G(M)}\}$. Our theory is stated for general K, but throughout this paper we use $K = 2$, which suffices empirically and keeps per-group calibration sets large.

3.2. Key Quantities and Coverage Bounds

Let $F_{\text{cal}}(t) = \mathbb{P}(S_i \le t)$, $F^k_{\text{cal}}(t) = \mathbb{P}(S_i \le t \mid G(M_i) = k)$, and $F^k_{\text{test}}(t) = \mathbb{P}(S_{n+1} \le t \mid G(M_{n+1}) = k)$ under the test distribution. Let $\pi_{\text{cal}}(k)$ and $\pi_{\text{test}}(k)$ be the group prevalences at calibration and test time.

Definition 3.1 (Calibration heterogeneity)**.** *For group $k \in [K]$,*

$$\eta_k := d_{\text{KS}}(F^k_{\text{cal}}, F_{\text{cal}}) = \sup_t \left| F^k_{\text{cal}}(t) - F_{\text{cal}}(t) \right|.$$

Intuitively, η_k measures how much pooled calibration distorts the score distribution for group k; it is nonzero whenever different groups have different score laws, even without any deployment shift.

Definition 3.2 (Within-group deployment shift)**.** *For group $k \in [K]$,*

$$\delta_k := d_{\text{KS}}(F^k_{\text{test}}, F^k_{\text{cal}}) = \sup_t \left| F^k_{\text{test}}(t) - F^k_{\text{cal}}(t) \right|.$$

Intuitively, δ_k captures the genuine deployment mismatch that remains even after conditioning on group membership; Mondrian calibration cannot eliminate this term.

Definition 3.3 (Between-group composition shift)**.**

$$\delta_{\text{btw}} := \tfrac{1}{2} \sum_{k=1}^{K} \left| \pi_{\text{test}}(k) - \pi_{\text{cal}}(k) \right|.$$

This measures how much the mixture of missingness-defined groups changes from calibration to deployment.

These three quantities play distinct roles. η_k is a *pooling artifact* addressable by conditioning. δ_k is genuine within-group deployment mismatch that conditioning cannot address. δ_{btw} captures composition shift in the group mixture and motivates the variable selection rule (§4).

Theorem 3.4 (Subgroup coverage bounds)**.** *Assume continuous score distributions and $n_k \ge \lceil \alpha^{-1} - 1 \rceil$.*

1. Standard CP*: for every group $k \in [K]$,*
$$\left| \mathbb{P}(Y_{n+1} \in C_{\text{std}} \mid G(M_{n+1}) = k) - (1-\alpha) \right| \le \eta_k + \delta_k + \tfrac{1}{n+1}.$$
2. Mondrian CP*: conditional on the realized calibration group assignments, for every group $k \in [K]$,*
$$\left| \mathbb{P}(Y_{n+1} \in C_{\text{mon}} \mid G(M_{n+1}) = k) - (1-\alpha) \right| \le \delta_k + \tfrac{1}{n_k+1}.$$
3. the Mondrian bound is tighter *whenever* $\eta_k > \frac{1}{n_k+1} - \frac{1}{n+1}$.

The proof applies a triangle-inequality decomposition: $|F^k_{\text{test}}(\hat{q}) - (1-\alpha)| \le |F^k_{\text{test}}(\hat{q}) - F^k_{\text{cal}}(\hat{q})| + |F^k_{\text{cal}}(\hat{q}) - F_{\text{cal}}(\hat{q})| + |F_{\text{cal}}(\hat{q}) - (1-\alpha)|$, bounded by δ_k, η_k, and $1/(n+1)$ respectively. For Mondrian CP the same argument applies within group k, so the η_k term does not appear in the bound. In the pure heterogeneity regime ($\delta_k = 0$), Mondrian is bounded by $1/(n_k+1)$ while standard CP still incurs up to $\eta_k + 1/(n+1)$, isolating what hard conditioning is designed to fix: pooling groups with different score laws, not within-group deployment shift.

3.3. Distribution Shift Decomposition and Portability

Lemma 3.5 (Mixture decomposition of distribution shift)**.**

$$d_{\mathrm{KS}}(F_{\mathrm{test}},\, F_{\mathrm{cal}}) \;\le\; \delta_{\mathrm{btw}} + \sum_{k=1}^{K} \pi_{\mathrm{test}}(k)\, \delta_k,$$

where

$$F_{\mathrm{cal}}(t) = \sum_k \pi_{\mathrm{cal}}(k) F^k_{\mathrm{cal}}(t), F_{\mathrm{test}}(t) = \sum_k \pi_{\mathrm{test}}(k) F^k_{\mathrm{test}}(t).$$

Cross-hospital shift therefore has at least two layers: composition shift in the prevalence of groups (δ_{btw}), and residual score shift within groups (δ_k). Lemma 3.5 motivates the variable selection rule in §4: choosing the missingness variable with the largest cross-site divergence targets the variable for which δ_{btw} is largest, aligning the grouping with the dominant composition-shift component.

Proposition 3.6 (Portability)**.** *Let $G_M : \{0,1\}^d \to [K]$ depend only on the missingness mask. For any two predictors $\hat{f}_1, \hat{f}_2$ and imputation maps Φ_1, Φ_2,*

$$G_M(M; \hat{f}_1, \Phi_1) = G_M(M; \hat{f}_2, \Phi_2) = G_M(M).$$

Hence mask-based subgroup assignment *is invariant to model updates. By contrast, risk-based grouping $G_R(X, M) = h(\hat{f}(\Phi(X, M)))$ is not: replacing $(\hat{f}, \Phi)$ can change a patient's subgroup even when their missingness pattern is unchanged.*

Corollary 3.7 (Scoped benefit of hard conditioning)**.** *When $\eta_k \gg 1/(n_k+1) - 1/(n+1)$, the Mondrian bound eliminates the dominant η_k term, providing a robust first-order reduction of the subgroup coverage gap. Conversely, when η_k is small (e.g., near-MCAR), the finite-sample grouping cost is non-negligible and hard conditioning need not improve upon pooled calibration.*

4. Method: Missingness-Aware Mondrian CP

The method has three components.

Four-way hospital-disjoint split. Hospitals: training (60%), selection (10%), calibration (10%), test (20%). The selection split is reserved solely for grouping-variable selection, preventing overfitting to calibration or test outcomes.

Label-free variable selection. For each feature j with missingness rate in $[10\%, 90\%]$ on the selection set:

$$j^* = \arg\max_{j \in \mathcal{V}} d_{\mathrm{KS}}\big(P_{\mathrm{sel}}(\mathbf{1}[X_j \text{ miss.}]),\; P_{\mathrm{cal}}(\mathbf{1}[X_j \text{ miss.}])\big). \tag{1}$$

This rule is *label-free*, *model-free*, and *test-blind*. It operationalizes Lemma 3.5: choosing the variable with the largest cross-site missingness divergence targets the grouping for which the between-group composition shift δ_{btw} is largest. The selected variable j^* varies across seeds (5 distinct variables over 10 seeds; App. J), but per-seed subgroup gap distributions overlap heavily, indicating the method's robustness derives from the selection rule rather than any specific biomarker.

Algorithm 1 Missingness-Aware Mondrian Conformal Prediction

Require: Model $\hat{f}$, splits $\mathcal{D}_{\mathrm{sel}}, \mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{te}}$, level α
Stage 1: Label-free variable selection
$\mathcal{V} \leftarrow \{j : 0.1 \le \text{miss-rate}(j, \mathcal{D}_{\mathrm{sel}}) \le 0.9\}$
$j^* \leftarrow \arg\max_{j \in \mathcal{V}} d_{\mathrm{KS}}\big(P_{\mathrm{sel}}(M_j),\, P_{\mathrm{cal}}(M_j)\big)$
$G(M) \leftarrow \mathbf{1}[M_{j^*} = 1]$
Stage 2: Mondrian calibration
for $g \in \{0, 1\}$ **do**
  $\mathcal{C}_g \leftarrow \{s_i : G(M_i) = g,\; i \in \mathcal{D}_{\mathrm{cal}}\}$
  if $|\mathcal{C}_g| \ge \lceil \alpha^{-1} - 1 \rceil$ **then**
    $\hat{q}_g \leftarrow \mathrm{Quantile}(\mathcal{C}_g,\, \lceil (1-\alpha)(|\mathcal{C}_g|+1) \rceil / |\mathcal{C}_g|)$
  else
    $\hat{q}_g \leftarrow \hat{q}_{\mathrm{pool}}$
  end if
end for
Stage 3: Prediction
Return $\Gamma(z) = \{y : s(\hat{f}(z), y) \le \hat{q}_{G(M)}\}$ for each $(z, M) \in \mathcal{D}_{\mathrm{te}}$

Mondrian calibration with fallback. The binary grouping $G(M) = \mathbf{1}[M_{j^*} = 1]$ yields two subgroups with separate thresholds. If a subgroup calibration set is smaller than $\lceil 1/\alpha - 1 \rceil$, we revert to the global threshold.

5. Experiments

5.1. Simulation Study

We simulate $d = 20$ features with severity-coupling γ controlling MCAR ($\gamma = 0$) vs. MAR ($\gamma > 0$). Figure 1 shows the key result: under MCAR, standard and Mondrian CP are indistinguishable ($\eta_k \approx 0$); under MAR-severity ($\gamma = 2$), Mondrian compresses the subgroup gap from 0.07–0.08 to 0.03–0.05. Per-hospital analysis of GOSSIS estimates $\gamma \approx 2.8$ (SD 0.63), placing it in the MAR regime where Corollary 3.7 predicts a benefit.

5.2. GOSSIS Hospital-Disjoint Benchmark

Setup. GOSSIS-1/eICU (Pollard et al., 2018): 106,488 admissions, 82 hospitals, 183 features, 9.2% mortality. Three base models: logistic regression, XGBoost (Chen & Guestrin, 2016), MLP. Results averaged over 10 hospital-disjoint seeds.

Main result (Figure 2). Standard CP achieves 90%

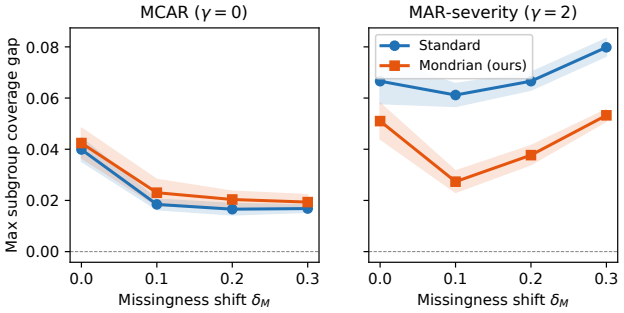

Figure 1. Max subgroup coverage gap vs. missingness shift δ_M (20 seeds, shaded ± 1 s.e.). **Left:** MCAR ($\gamma = 0$): methods are indistinguishable. **Right:** MAR-severity ($\gamma = 2$): standard CP degrades while Mondrian remains stable.

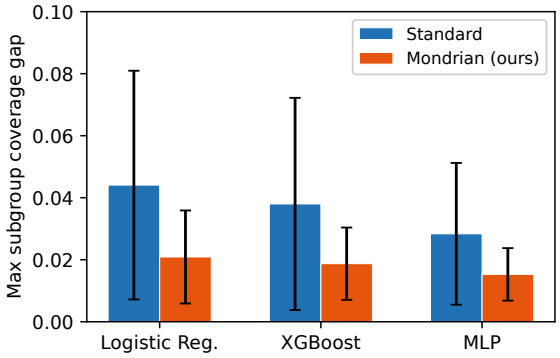

Figure 2. Max subgroup coverage gap on GOSSIS across three base models (mean $\pm$ std, 10 seeds). Mondrian CP compresses the gap by $\approx$50% consistently.

marginal coverage but leaves subgroup gaps of 0.028–0.044. Mondrian CP compresses these to 0.015–0.021 ($\approx$50% reduction) with $<$1% set-size change. Across all 891 variable–seed combinations, estimated η_k (mean 0.054) far exceeds the grouping cost $1/(n_k{+}1) - 1/(n{+}1) \approx 10^{-4}$, confirming Theorem 3.4's advantage condition.

Baselines (Table 1). Predicted-risk grouping attains the lowest gap on the fixed benchmark (0.017 vs. 0.019) but has portability 0.20: 80% of patients change subgroup under model updates, consistent with Proposition 3.6. Gibbs (missingness features) shares Portability = 1.000 with our method because both condition on the mask; the distinction is robustness under shift. Optimization-based baselines (Gibbs (2025), learned partition (2024)) are competitive in-distribution but degrade under MIMIC-IV deployment stress.

5.3. MIMIC-IV Care-Unit-Disjoint Validation

Table 2 shows independent validation on MIMIC-IV (Johnson et al., 2023) (85,181 admissions, 6 care units, leave-one-unit-out, 18 splits). The same pattern holds across all three base models.

Table 1. Grouping comparison on GOSSIS (XGBoost, 10 seeds). Portability = fraction of patients with identical subgroup assignment across 3 base models.

Method	Max gap	Set size	Portability
Mondrian, missingness (ours)	$.019 \pm .012$	$.960 \pm .014$	**1.000**
Mondrian, predicted risk	$\mathbf{.017 \pm .008}$	$.961 \pm .009$	.202
Mondrian, APACHE severity	$.018 \pm .009$	$.960 \pm .010$	—
Gibbs (2025), miss. features	$.033 \pm .016$	$1.025 \pm .017$	1.000
Learned partition (2024)	$.031 \pm .013$	$1.004 \pm .022$	—
Weighted CP (2019)	$.042 \pm .041$	$.956 \pm .012$	1.000
Standard (no grouping)	$.038 \pm .034$	$.959 \pm .011$	—

Table 2. MIMIC-IV care-unit-disjoint validation (mean $\pm$ std, 18 splits).

Model	Method	Max gap	Coverage
Logistic	Standard	$.075 \pm .032$	$.883 \pm .041$
	Mondrian (ours)	$\mathbf{.046 \pm .023}$	$.889 \pm .037$
XGBoost	Standard	$.054 \pm .022$	$.904 \pm .038$
	Mondrian (ours)	$\mathbf{.037 \pm .026}$	$.907 \pm .032$
MLP	Standard	$.049 \pm .015$	$.903 \pm .032$
	Mondrian (ours)	$\mathbf{.026 \pm .016}$	$.909 \pm .024$

Deployment stress (App. D). MAR-style dropout on the MIMIC-IV test set simulates a site ordering fewer informative labs. At drop rate 0.5, the Gibbs missingness baseline degrades severely (coverage 0.571, gap 0.373), while our method remains stable (0.917, gap 0.028). The learned partition is more robust (0.923, gap 0.049) but requires supervised fitting. Hard conditioning outperforms more expressive baselines when shift is structural.

6. Discussion

The method's benefit is scoped: it is most effective when η_k dominates the grouping cost $1/(n_k{+}1) - 1/(n{+}1)$ (Theorem 3.4, part iii), and neutral under near-MCAR regimes where $\eta_k \approx 0$. The residual gap reflects within-group shift δ_k, which conditioning cannot remove (part ii); combining groupwise calibration with methods that target δ_k is a natural next step.

Portability is a key operational advantage: once j^* is selected, subgroup assignment never changes under model retraining (Proposition 3.6). The method requires no retraining, no protected-attribute access, and negligible computation. Patients whose data are systematically less complete should not receive less reliable uncertainty estimates; this work offers a theoretically grounded way to reduce that gap under cross-hospital deployment.

References

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the*

IMA, 10(2):455–482, 2021.

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

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Fan, J. et al. Weighted conformal prediction provides adaptive and valid mask-conditional coverage for general missing data mechanisms. *arXiv preprint arXiv:2512.14221*, 2024.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B*, 87(4):1100–1126, 2025.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Liu, Y., Hu, K., Zou, Z., Tian, J., and Lei, J. Multi-source conformal inference under distribution shift. In *International Conference on Machine Learning (ICML)*, 2024.

Martínez Gil, A. et al. Identifying homogeneous and interpretable groups for conformal prediction. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, pp. 345–356. Springer, 2002.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

Rockenschaub, P. et al. Robust prediction under missingness shifts. *arXiv preprint arXiv:2406.16484*, 2024.

Romano, Y., Patterson, E., and Candès, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sharafoddini, A., Dubin, J. A., Maslove, D. M., and Lee, J. A new insight into missing data in intensive care unit patient profiles: Observational study. *JMIR Medical Informatics*, 7(1):e11605, 2019.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Tipirneni, S. and Reddy, C. K. On missingness features in machine learning models for critical care: Observational study. *JMIR Medical Informatics*, 9(12):e25022, 2021.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005.

Zaffran, M., Dieuleveut, A., Josse, J., and Romano, Y. Conformal prediction with missing values. In *International Conference on Machine Learning (ICML)*, pp. 40578–40604, 2023.

A. Proof of Theorem 3.4

Proof. **Part (i).** Since $\{S_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} F_{\text{cal}}$ and F_{cal} is continuous, $U_i := F_{\text{cal}}(S_i) \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$, so $F_{\text{cal}}(\hat{q}) = U_{(r)}$ and $\mathbb{E}[F_{\text{cal}}(\hat{q})] = r/(n+1) \in [1-\alpha,\, 1-\alpha+1/(n+1)]$ (Romano et al., 2019, Lemma 1). By independence of test and calibration:

$$\begin{aligned}&|\mathbb{P}(S_{n+1} \le \hat{q} \mid G_{n+1}{=}k) - (1{-}\alpha)| \\ &\le \mathbb{E}|F^k_{\text{test}}(\hat{q}) - F^k_{\text{cal}}(\hat{q})| + \mathbb{E}|F^k_{\text{cal}}(\hat{q}) - F_{\text{cal}}(\hat{q})| \\ &\quad + |\mathbb{E}[F_{\text{cal}}(\hat{q})] - (1{-}\alpha)| \;\le\; \delta_k + \eta_k + \tfrac{1}{n+1}. \quad (2)\end{aligned}$$

The pointwise bounds $|F^k_{\text{test}}(t) - F^k_{\text{cal}}(t)| \le \delta_k$ and $|F^k_{\text{cal}}(t) - F_{\text{cal}}(t)| \le \eta_k$ hold for every realization of $\hat{q}$.

Part (ii). Since $g_i = G(M_i)$ depends only on M_i, conditioning on $g_{1:n}$ preserves within-group i.i.d. structure. The within-group scores $\{S_i : i \in \mathcal{C}_k\}$ remain i.i.d. from F^k_{cal}, giving $\mathbb{E}[F^k_{\text{cal}}(\hat{q}_k) \mid g_{1:n}] \in [1-\alpha,\, 1-\alpha+1/(n_k+1)]$. The bound follows Part (i) with η_k absent.

Part (iii). Subtracting bounds directly: $(\eta_k + \delta_k + \frac{1}{n+1}) - (\delta_k + \frac{1}{n_k+1}) = \eta_k - (\frac{1}{n_k+1} - \frac{1}{n+1})$. □

B. Mixture decomposition lemma

Lemma B.1. $d_{\text{KS}}(F_{\text{test}}, F_{\text{cal}}) \le \delta_{\text{btw}} + \sum_k \pi_{\text{test}}(k)\, \delta_k$.

Proof sketch. Add and subtract $\pi_{\text{test}}(k) F^k_{\text{cal}}(t)$ inside the KS supremum. The within-group residual is bounded by $\sum_k \pi_{\text{test}}(k)\delta_k$; the between-group mixture term is bounded by δ_{btw} via a vertex argument on the simplex. □

C. Dataset details

GOSSIS features: 183 raw-derived variables after removing identifiers, outcomes, leakage columns, death-probability scores, all-NaN columns, and constants. MIMIC-IV: 160

Table 3. Dataset Summary Statistics.

Dataset	Cohort	Patients	Sites	Mortality	Mean miss.	Range
GOSSIS	≥500 adm.	106,488	82	9.2%	.332	[.00, .92]
MIMIC-IV	≥10K adm.	85,181	6	11.7%	.372	[.35, .40]

first-day ICU features (demographics, vitals, labs, blood gas, urine output, GCS/SOFA); care units serve as pseudo-hospitals, inducing compound missingness and case-mix shift.

D. MIMIC-IV missingness stress test

To test whether the advantage scales with missingness heterogeneity, we apply MAR-style dropout to the MIMIC-IV test set: for each split, the top-5 predictive features (by XGBoost importance) are independently masked with probability δ_{drop}, simulating a site that orders fewer informative labs. Standard CP's gap increases from 0.054 to 0.063 as δ_{drop} grows from 0 to 0.5, while missingness-aware Mondrian CP remains lower at each point (0.037 to 0.061). The advantage is clearest at moderate perturbation ($\delta_{\text{drop}} \leq 0.3$); at extreme dropout the two methods converge as the perturbation overwhelms the calibration structure.

The stress test also reveals the most striking failure among baselines. Gibbs with missingness features (Gibbs et al., 2025), which achieves the lowest gap on GOSSIS (0.027; Table 1), degrades sharply under stress: coverage falls to ∼0.57–0.59 across all drop rates, with gap 0.36–0.38. This occurs because the optimized threshold function $\hat{q}(x) = \langle \hat{\theta}, \phi(x) \rangle$ is fitted to the calibration distribution's missingness structure and extrapolates poorly when the test distribution shifts. By contrast, Gibbs with general features remains stable (coverage ∼0.92–0.94), as does the learned partition of Martínez Gil et al. (2024) (coverage ∼0.92). Our method maintains coverage 0.92 throughout. This demonstrates that the failure is specific to *optimizing over missingness features under shift*, not to the Gibbs framework itself; simple Mondrian conditioning provides implicit regularization against this failure mode.

E. Model-swap analysis

Under model swap (calibrate with model A, deploy model B), stale thresholds degrade all methods, but not equally. Missingness grouping preserves subgroup structure exactly (self-overlap 1.0), whereas risk-based grouping changes substantially after a model update. In the real-data swap matrix, missingness grouping yields a lower subgroup gap than frozen-risk grouping in 50/90 swap cells and than matched-risk grouping in 67/90 cells. Controlled simulation (Sim 5) shows the same pattern under MAR-severity: swapped missingness grouping remains better than swapped risk grouping even under severe model mismatch.

F. Calibration set size sensitivity

Missingness-aware Mondrian CP improves over standard across all calibration sizes (500 to full). The finite-sample cost drops from 0.006 at $|D_{\text{cal}}| = 500$ to 0.0003 at full size, while η_k remains stable, consistent with Theorem 3.4.

G. Grouping ablations

Multi-variable composite grouping (top-3 by KS distance) achieves comparable gap compression but lower stability (Jaccard 0.10 vs. 0.24 for single-variable). Increasing to $K > 2$ groups (e.g., terciles of the selected variable) was not explored because the binary grouping already provides ∼10,000 patients per group on GOSSIS; further splitting would reduce per-group calibration size without addressing the within-group shift δ_k that drives the residual gap. The shrinkage parameter of the weighted within-group variant has negligible effect: shrinkage $\in \{0, 0.5, 0.7, 1.0\}$ all yield gap ∼0.017 (5-seed repeated). Pure Mondrian (shrinkage $= 0$) is adopted as the default.

H. Sensitivity to coverage level

At $\alpha = 0.05$ (95% target coverage, XGBoost, GOSSIS, 10 seeds), the gap compression pattern is preserved: standard gap 0.015 ± 0.012 vs. Mondrian gap 0.010 ± 0.006. Set sizes increase to ∼1.07 as expected for a more conservative target. The relative improvement is consistent with the $\alpha = 0.10$ results, confirming that the method is not sensitive to the choice of coverage level.

I. Simulation details

- **Sim 1:** $\delta_M \in \{0, 0.1, 0.2, 0.3\}$, $\gamma \in \{0, 1, 2, 3\}$, 10 seeds; logistic regression with $n_{\text{cal}} = 1000$ and $n_{\text{test}} = 1500$.
- **Sim 2:** $\delta_M \times \delta_X$ grid (7×5), $\gamma \in \{0, 2\}$, 20 seeds.
- **Sim 3:** $\delta_M = 0.2$, MCAR/MAR-severity/MAR-selective/MNAR, 20 seeds.
- **Sim 4:** Per-hospital missingness–severity correlation with mean $\gamma \approx 2.8$ and SD 0.63.
- **Sim 5:** Model swap (logistic→XGBoost), 20 seeds.

J. Variable selection stability

Across 10 seeds (identical for all three base models), variable selection frequencies are shown below. All selected variables are clinically relevant, and grouping is identical across models for every seed.

Variable	Freq. (10 seeds)
urineoutput_apache	5/10
h1_glucose_max	2/10
d1_lactate_max	1/10
d1_pao2fio2ratio_max	1/10
hospital_admit_source	1/10

Downstream robustness. We stratify the 10-seed missingness-aware gap by selected variable identity. The resulting gap distributions overlap heavily, indicating that the selection rule remains robust even when different variables are selected.

Table 4. Subgroup Gap by Selected Variable (Missingness-Aware Mondrian CP, 10 Seeds).

Model	Variable	Seeds	Gap (mean)	Gap (range)
Logistic reg.	urineoutput_apache	5	.025	[.007, .043]
	other variables	5	.017	[.002, .040]
XGBoost	urineoutput_apache	5	.023	[.016, .042]
	other variables	5	.015	[.002, .031]
MLP	urineoutput_apache	5	.015	[.006, .032]
	other variables	5	.015	[.004, .023]

K. Kernel-smoothed CP exploration

We implement kernel-smoothed CP with Hamming kernel on the top-5 missingness indicators selected by our label-free rule, with bandwidth h chosen on $\mathcal{D}_{\text{sel}}$ from a grid $h \in \{0, 0.5, 1, 2, 5, \infty\}$ by minimizing the selection-split subgroup gap. On GOSSIS (XGBoost, 10 seeds), the selected h^* distribution is: $h = 0.5$ in 4/10 seeds, $h = 1.0$ in 2/10, $h = 0$ in 2/10, $h = \infty$ in 2/10. The mean gap is 0.020 ± 0.013, compared to 0.019 ± 0.012 for binary Mondrian and 0.038 ± 0.034 for standard CP. The non-trivial h^* in 6/10 seeds confirms that the continuum between standard and Mondrian CP is empirically meaningful, but hard conditioning (binary Mondrian) is near-optimal in this setting.

L. Selection split ratio ablation

We vary the selection/calibration split while keeping train (60%) and test (20%) fixed. With only 5% of hospitals for selection (sel/cal = 5/15), variable selection becomes noisier (7 unique variables across 10 seeds vs. 5 at the default 10/10), and Mondrian gap degrades on individual seeds (e.g., gap $= 0.131$ when hospital_bed_size is selected). At the default 10/10 split, Mondrian gap is 0.025 ± 0.017. A larger selection split (15/5) yields 6 unique variables and gap 0.027 ± 0.012, but the smaller calibration set increases threshold variance. The 10/10 default balances selection stability against calibration precision.

M. Coverage asymmetry decomposition

We decompose the symmetric max gap into worst under-coverage ($\max_k[(1-\alpha) - \text{coverage}_k]^+$) and worst over-coverage ($\max_k[\text{coverage}_k - (1-\alpha)]^+$) across both datasets and all three base models.

Table 5. Coverage Asymmetry: Worst Under- and Over-Coverage (Mean $\pm$ Std).

Dataset	Model	Worst under-cov.		Worst over-cov.	
		Standard	Mondrian	Standard	Mondrian
GOSSIS	Logistic reg.	$.035 \pm .042$	$.011 \pm .016$	$.021 \pm .021$	$.011 \pm .014$
	XGBoost	$.031 \pm .038$	$.011 \pm .011$	$.018 \pm .019$	$.011 \pm .013$
	MLP	$.023 \pm .025$	$.010 \pm .007$	$.014 \pm .018$	$.008 \pm .011$
MIMIC-IV	Logistic reg.	$.064 \pm .045$	$.033 \pm .031$	$.021 \pm .021$	$.015 \pm .020$
	XGBoost	$.038 \pm .032$	$.018 \pm .028$	$.027 \pm .024$	$.021 \pm .024$
	MLP	$.036 \pm .028$	$.010 \pm .016$	$.025 \pm .019$	$.016 \pm .018$

Mondrian CP improves both sides, but the largest absolute improvement is consistently on the under-coverage side. On GOSSIS with XGBoost, worst under-coverage drops from 0.031 to 0.011 (65% reduction), while worst over-coverage drops from 0.018 to 0.011 (39%). Since under-coverage means patients receive unreliably optimistic prediction sets, this asymmetry is clinically favorable.

N. Score Design Ablation

Our method uses missingness only in the threshold layer through Mondrian conditioning and leaves the conformity score unchanged. A natural follow-up is whether one can do better by also making the score itself missingness-aware. We evaluate three score-design families and find a consistent pattern. On GOSSIS, score redesign can modestly improve subgroup gap, but the gain beyond Mondrian thresholding is very small. On MIMIC-IV, the only variant with consistent benefit is the simplest one: global-missingness scaling combined with Mondrian conditioning. More aggressive transformations can substantially change the score distribution, and even drive the calibration heterogeneity term η_k close to zero, without materially reducing the residual subgroup gap.

Score variants. We evaluate three modifications of the base score $s_i = 1 - \hat{p}_i[y_i]$:

- **Missingness-scaled score.** $\tilde{s}_i = s_i \cdot (1 + \beta \cdot m_i)$, where m_i is either the global missingness rate or the selected variable's binary indicator. We choose β on $\mathcal{D}_{\text{sel}}$ from $\{0, 0.5, 1, 2, 5, 10\}$ to minimize subgroup gap.
- **Z-score normalization.** $\tilde{s}_i = (s_i - \mu_{g(i)})/\sigma_{g(i)}$, with location-only, location-scale, and quantile-normalization variants estimated from the calibration split.
- **Linear recalibration.** $\tilde{s}_i = s_i - \hat{h}(m_i)$, where $\hat{h}$ is a linear model fitted on $\mathcal{D}_{\text{sel}}$ using either the selected variable or the top-5 missingness indicators ranked by KS

shift.

All three families use selection or calibration labels, so unlike the default method they are not label-free.

Results on GOSSIS. Table 6 reports the main GOSSIS comparison (XGBoost, 10 seeds). Three points matter. First, score adjustment can help under standard CP: the best score-only variant, selected-variable scaling, reduces the gap from 0.040 to 0.023. Second, once Mondrian conditioning is applied, the additional gain is small. Original Mondrian attains gap 0.0246, and the best adjusted-score Mondrian variant (global-missingness scaling) attains 0.0240. Third, several groupwise transformations are effectively invariant under Mondrian thresholding. In particular, selected-variable scaling, selected-variable recalibration, and location or location-scale normalization yield the same GOSSIS Mondrian numbers as the original score, because they preserve within-group score ordering and therefore do not change the groupwise conformal quantile.

Table 6. Score design ablation on GOSSIS, XGBoost (mean over 10 seeds). Each row is a score variant; columns compare standard and Mondrian thresholding. Original Mondrian (row 1, right) is the proposed method.

	Standard CP			Mondrian CP		
Score variant	Cov.	Gap	Size	Cov.	Gap	Size
Original	.895	.040	.952	.894	.025	.953
Scaled (global m)	.896	.033	.958	.894	.024	.959
Scaled (selected var.)	.894	.023	.953	.894	.025	.953
Z-norm (location)	.895	.034	.951	.894	.025	.953
Z-norm (loc-scale)	.895	.032	.951	.894	.025	.953
Z-norm (quantile)	.894	.025	.953	.894	.024	.954
Recalib. (selected var.)	.894	.033	.951	.894	.025	.953
Recalib. (top-5 KS)	.895	.034	.951	.894	.024	.953

Heterogeneity decomposition. Table 7 shows why these gains saturate. Quantile normalization drives η_k from 0.0706 to 7×10^{-4}, but the Mondrian gap remains 0.0244 and δ_k is essentially unchanged (0.0475 to 0.0474). The remaining error is therefore not pooled-versus-group mismatch, which Mondrian already removes, but residual within-group shift. The table also shows the opposite failure mode: top-5 recalibration does not help because it perturbs both η_k and δ_k in the wrong direction, increasing residual instability rather than reducing it.

Robustness under deployment shift (MIMIC-IV). The GOSSIS result alone would be too optimistic. On MIMIC-IV leave-one-unit-out validation and the stress test, the only score-design variant with consistent improvement is global-missingness scaling plus Mondrian conditioning. It improves the LOO gap from 0.129 to 0.119 and the stress gap at drop rate 0.5 from 0.129 to 0.120 (Table 8). By contrast, selected-variable scaling does not move the Mondrian result at all, and more aggressive adjustments can be brittle: top-5 recalibration under Mondrian is roughly neutral on LOO (gap 0.130) but degrades under stress (gap 0.139, coverage 0.820). This pattern mirrors the broader paper story. Score redesign can be helpful when it is simple and aligned with the deployment shift, but richer calibration-fitted transformations are easier to destabilize under shift.

Table 7. Calibration heterogeneity and residual shift under different score designs on GOSSIS (XGBoost, 10 seeds). Eliminating η_k does not eliminate the Mondrian subgroup gap because the residual term is dominated by δ_k.

Score variant	η_k (mean)	η_k (max)	δ_k (mean)	Mondrian gap
Original	.0706	.276	.0475	.0246
Scaled (global m)	.0621	.210	.0462	.0240
Z-norm (loc-scale)	.1344	.350	.0475	.0246
Z-norm (quantile)	.0007	.002	.0474	.0244
Recalib. (top-5 KS)	.1092	.352	.0655	.0242

Table 8. MIMIC-IV score-design validation (XGBoost). LOO columns report leave-one-unit-out means; stress columns report the hardest setting with drop rate 0.5. The only variant with consistent benefit across both settings is global-missingness scaling with Mondrian thresholding.

Method	LOO cov.	LOO gap	Stress cov.	Stress gap
Original standard	.826	.149	.826	.148
Original Mondrian	.835	.129	.835	.129
Scaled (global m) + standard	.834	.129	.832	.131
Scaled (global m) + Mondrian	.839	.119	.837	.120
Z-norm (loc-scale) + standard	.830	.123	.831	.122
Recalib. (top-5 KS) + Mondrian	.833	.130	.820	.139

Takeaway. Score-level missingness adjustment and threshold-level Mondrian conditioning act on different parts of the decomposition in Theorem 3.4. Score redesign can change η_k and sometimes slightly improve the final gap, but the residual post-Mondrian error is driven mainly by δ_k. That is why even strong distributional reshaping of the score produces little extra subgroup-coverage benefit. The simplest consistent refinement is global-missingness scaling plus Mondrian conditioning; more elaborate variants are usually neutral or fragile under deployment shift. This is why the main method keeps the score unchanged and concentrates missingness information in the threshold layer, preserving portability and the label-free grouping rule while already addressing the dominant pooled-heterogeneity term.