TAIL DEPENDENCE, NOT INTRACLASS CORRELATION: DESIGN EFFECTS AND A CLUSTER FLOOR FOR CONFORMAL CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction is often calibrated on clustered data, such as patients within hospitals or cells within donors. Clustering has little effect on marginal coverage, but it can substantially change training-conditional coverage, the coverage induced by the calibration set at hand. We give an exact finite-sample characterization of this effect. The calibration counting process satisfies a design-effect identity governed by the within-cluster correlation of threshold indicators rather than by the intraclass correlation of the scores. At extreme thresholds, this indicator correlation converges to the upper tail-dependence coefficient of the within-cluster copula. Thus two problems with the same intraclass correlation can require very different corrections: the design-effect penalty vanishes for Gaussian dependence as the miscoverage level shrinks, whereas it persists for a tail-dependent copula. The identity connects Vovk's i.i.d. Beta law to the comonotone regime, where the effective sample size is the number of clusters. We also prove a cluster floor: no rule can guarantee PAC coverage, and no procedure can guarantee conformal risk control, from fewer than $\log \delta / \log(1-\alpha)$ clusters; for $\alpha = \delta = 0.1$, this is 22 clusters. Finally, we construct dependence for which a single-number summary of within-cluster correlation understates the design effect by a factor approaching the cluster size. On multi-site neuroimaging, threshold-level calibration is the only implementable method among those compared that meets the empirical PAC target.

1 INTRODUCTION

Split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018) turns any point predictor into a set-valued predictor with distribution-free marginal coverage, provided calibration and test data are exchangeable. The guarantee now extends to general monotone losses (Angelopoulos et al., 2024), to covariate shift (Tibshirani et al., 2019), and to conditional statements over a chosen shift class (Gibbs et al., 2025); Angelopoulos et al. (2025a) give a book-length treatment. Its practical appeal in scientific and clinical machine learning is that it places little structure on the predictive model. Its central assumption, exchangeability, often fails in these settings because observations arrive in *clusters*. Subjects are scanned at one of a handful of imaging sites, cells are sequenced from one of a few dozen donors, and measurements repeat within a patient. Observations that share a scanner, protocol, reagent batch or cohort are therefore dependent. Cluster-randomised trials are designed around this structure (Donner & Klar, 2000), and multi-site neuroimaging studies document its effect on reproducibility (Marek et al., 2022; Rosenblatt et al., 2024).

Clustering is usually discussed through marginal coverage, which can remain nearly unchanged. For a fixed calibration sample, the more informative quantity is the *training-conditional* coverage

$$C \;=\; \mathbb{P}\big(Y_{\text{test}} \in \widehat{C}(X_{\text{test}}) \,\big|\, \mathcal{D}_{\text{cal}}\big), \tag{1}$$

the coverage induced by the calibration set actually used in deployment. Marginal coverage is $\mathbb{E}[C]$, so a procedure can meet its marginal target even when C falls below target for many calibration draws. In a recent multi-institution brain-tumour study, pooled calibration met its average target while violating it at 40% of individual sites (Shahid, 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

For i.i.d. calibration data the training-conditional law is known exactly, $C \sim \mathrm{Beta}(k, n+1-k)$ (Vovk, 2013), and it underlies the standard PAC correction and the analysis of Bian & Barber (2023). Under clustering, results exist at two extremes. Dunn et al. (2023) obtain finite-sample marginal validity by keeping one observation per cluster, which concedes that the effective sample size is the number of clusters. Lee et al. (2025) give marginal coverage under hierarchical exchangeability, with slack governed by the number of groups. Most recently, Ramos et al. (2026) transport Vovk's law to a clustered model in which all scores within a cluster coincide, so that the effective sample size is exactly the number of clusters. Real clustered datasets lie between these endpoints. We characterize that intermediate regime.

Contributions.

1. **Coverage as an order statistic (Prop. 1).** For a test point from an unseen cluster, training-conditional coverage equals the k-th order statistic of the probability-integral transforms of the calibration scores, so clustering acts only through their joint law.
2. **An exact design-effect identity (Thm. 1).** $\mathrm{Var}[N(t)] = n\,t(1-t)[1 + (\bar{m}_a - 1)\rho_{\mathrm{ind}}(t)]$ holds exactly for any within-cluster exchangeable dependence and for unequal cluster sizes. The parameter is the correlation of the threshold *indicator*; substituting the score correlation overstates the variance by 37 to 91% in our simulations, with the error growing into the tail.
3. **Tail dependence governs the correction (Thm. 2).** $\rho_{\mathrm{ind}}(t) = [\lambda_{\mathrm{U}}(t) - (1-t)]/t \to \lambda_{\mathrm{U}}$ as $t \to 1$, where λ_{U} is the upper tail-dependence coefficient of the within-cluster copula. Gaussian dependence ($\lambda_{\mathrm{U}} = 0$) becomes harmless as $\alpha \to 0$; Student-t dependence ($\lambda_{\mathrm{U}} > 0$) does not. Vovk's law and the comonotone regime of Ramos et al. (2026) are the endpoints $\rho_{\mathrm{ind}} \equiv 0$ and $\lambda_{\mathrm{U}} = 1$.
4. **A cluster floor (Prop. 2, Cor. 4).** Distribution-free PAC coverage and distribution-free risk control both require at least $\log\delta / \log(1-\alpha)$ clusters.
5. **The design effect is not a scalar (Thm. 3).** We construct dependence for which the design effect is small below a threshold but equals the cluster size above it. A one-number summary of dependence can then under-correct by a factor approaching $\bar{m}_a$.
6. **A threshold-level correction (§5),** evaluated on copula simulations, on ABIDE (24 imaging sites), and on a 356,213-cell atlas grouped by 44 donors.

2 Setup

Data. Calibration data comprise b clusters, cluster j contributing m_j observations, $n = \sum_{j=1}^{b} m_j$. Write the conformity scores as $S_{j,r}$, $r = 1, \ldots, m_j$. We assume:

(A1) *Common marginal.* Each $S_{j,r}$ has the same continuous marginal CDF F.

(A2) *Within-cluster exchangeability.* For each j, the vector $(S_{j,1}, \ldots, S_{j,m_j})$ is exchangeable.

(A3) *Between-cluster independence.* The clusters are mutually independent.

(A4) *Fresh test cluster.* The test point comes from a cluster that is independent of, and absent from, the calibration data.

Assumption (A4) is the deployment case: calibrate on the sites available and deploy at a new one. Appendix D treats a test point drawn from a calibration cluster.

Probability integral transforms. Let $U_{j,r} = F(S_{j,r})$. By (A1) each $U_{j,r}$ is marginally $\mathrm{Unif}(0,1)$, and by (A2)–(A3) the collection is clustered. Define the calibration counting process

$$N(t) \;=\; \sum_{j=1}^{b} \sum_{r=1}^{m_j} \mathbf{1}\{U_{j,r} \le t\}, \qquad t \in [0,1]. \tag{2}$$

Procedure. Split conformal with calibration index k thresholds at the k-th smallest calibration score $\widehat{s} = S_{(k)}$ and returns $\widehat{C}(x) = \{y : S(x, y) \leq \widehat{s}\}$. Nothing below depends on the choice of score, which may be a residual, a quantile-regression score (Romano et al., 2019), or any other conformity measure.

Definition 1 (Design effect)**.** For cluster sizes $m_1, \ldots, m_b$, the Kish mean cluster size (Kish, 1965) is $\bar{m}_a = \sum_j m_j^2 / \sum_j m_j$, and the design effect at threshold t is $\mathrm{DEFF}(t) = 1 + (\bar{m}_a - 1)\rho_{\mathrm{ind}}(t)$, where

$$\rho_{\mathrm{ind}}(t) \;=\; \mathrm{Corr}\left(\mathbf{1}\{U_{j,r} \leq t\},\, \mathbf{1}\{U_{j,r'} \leq t\}\right), \qquad r \neq r'. \tag{3}$$

3 Exact finite-sample theory

Proposition 1 (Coverage is an order statistic)**.** *Under (A1)–(A4), the training-conditional coverage of split conformal with calibration index k is*

$$C \;=\; F(S_{(k)}) \;=\; U_{(k)}, \tag{4}$$

the k-th order statistic of $\{U_{j,r}\}$. Consequently, for every t,

$$\mathbb{P}(C \leq t) \;=\; \mathbb{P}\big(N(t) \geq k\big). \tag{5}$$

Proof. By (A4) the test score is independent of the calibration set and has CDF F, so $C = \mathbb{P}(S_{\mathrm{test}} \leq \widehat{s} \mid \mathcal{D}_{\mathrm{cal}}) = F(\widehat{s})$. Since F is continuous and nondecreasing, $F(S_{(k)}) = U_{(k)}$. Equation (5) holds because $U_{(k)} \leq t$ exactly when at least k of the $U_{j,r}$ lie below t. □

The distribution of coverage is therefore the distribution of a sum of b independent within-cluster counts, and it equals $\mathrm{Beta}(k, n + 1 - k)$ only when the $U_{j,r}$ are i.i.d.

Theorem 1 (Exact design-effect identity)**.** *Under (A1)–(A3), for every $t \in (0, 1)$,*

$$\mathbb{E}[N(t)] = n\,t, \qquad \mathrm{Var}[N(t)] \;=\; n\,t(1-t)\,\big[\,1 + (\bar{m}_a - 1)\,\rho_{\mathrm{ind}}(t)\,\big]. \tag{6}$$

The identity is exact, requires no distributional assumption beyond (A1)–(A3), and holds for unequal cluster sizes.

Proof. Write $I_{j,r} = \mathbf{1}\{U_{j,r} \leq t\}$, so $\mathbb{E}I_{j,r} = t$ and $\mathrm{Var}\, I_{j,r} = t(1-t)$. By (A2), $\mathrm{Cov}(I_{j,r}, I_{j,r'}) = \rho_{\mathrm{ind}}(t)\, t(1-t)$ for $r \neq r'$, and by (A3) cross-cluster covariances vanish. Hence

$$\begin{aligned}\mathrm{Var}[N(t)] &= \sum_{j=1}^{b} \Big[m_j\, t(1-t) + m_j(m_j - 1)\,\rho_{\mathrm{ind}}(t)\, t(1-t)\Big] \\ &= t(1-t)\Big[n + \rho_{\mathrm{ind}}(t)\Big(\sum_j m_j^2 - n\Big)\Big] = n\,t(1-t)\Big[1 + \rho_{\mathrm{ind}}(t)\Big(\tfrac{\sum_j m_j^2}{n} - 1\Big)\Big], \end{aligned} \tag{7}$$

and $\sum_j m_j^2 / n = \bar{m}_a$ by definition. □

Remark 1 (Indicator versus score correlation)**.** The survey-sampling formula $\mathrm{DEFF} = 1 + (m - 1)\rho$ with ρ the correlation of the scores themselves describes the variance of a *mean*. Conformal calibration concerns an *order statistic*, and the relevant functional is the within-cluster concordance of the threshold indicators. The two differ even for Gaussian scores: at $t = 1/2$ Sheppard's formula gives $\rho_{\mathrm{ind}}(1/2) = (2/\pi)\arcsin \rho_S$, which is below ρ_S for $0 < \rho_S < 1$, and the gap widens in the tail where the $(1 - \alpha)$ quantile lies (Table 1).

Theorem 2 (Tail limit)**.** *Let $\lambda_{\mathrm{U}}(t) = \mathbb{P}(U_{j,r'} > t \mid U_{j,r} > t)$ for $r \neq r'$, and let $\lambda_{\mathrm{U}} = \lim_{t \to 1} \lambda_{\mathrm{U}}(t)$ be the upper tail-dependence coefficient of the within-cluster copula (Joe, 2014). Then for every $t \in (0, 1)$,*

$$\rho_{\mathrm{ind}}(t) \;=\; \frac{\lambda_{\mathrm{U}}(t) - (1-t)}{t}, \qquad \textit{hence} \qquad \lim_{t \to 1} \rho_{\mathrm{ind}}(t) \;=\; \lambda_{\mathrm{U}}. \tag{8}$$

Proof. Write $\bar{t} = 1-t$. By inclusion and exclusion, $\mathbb{P}(U \le t, U' \le t) = 1-2\bar{t}+\mathbb{P}(U > t, U' > t)$, and $\mathbb{P}(U > t, U' > t) = \bar{t}\,\lambda_{\mathrm{U}}(t)$ by the definition of $\lambda_{\mathrm{U}}(t)$. Since $t^2 = 1 - 2\bar{t} + \bar{t}^{\,2}$,

$$\mathbb{P}(U \le t, U' \le t) - t^2 = \bar{t}\,\lambda_{\mathrm{U}}(t) - \bar{t}^{\,2} = \bar{t}\big[\lambda_{\mathrm{U}}(t) - \bar{t}\big]. \tag{9}$$

Dividing by $\mathrm{Var}(\mathbf{1}\{U \le t\}) = t\bar{t}$ gives $\rho_{\mathrm{ind}}(t) = [\lambda_{\mathrm{U}}(t) - \bar{t}]/t$, and letting $t \to 1$ gives λ_{U}. □

Theorem 2 makes the design effect at the conformal threshold a tail property of the within-cluster copula. Three consequences follow.

Corollary 1 (Gaussian dependence)**.** *If the within-cluster copula is Gaussian with equicorrelation $\rho_S < 1$, then*

$$\rho_{\mathrm{ind}}(t) = \frac{\Phi_2\big(z_t, z_t; \rho_S\big) - t^2}{t(1-t)}, \qquad z_t = \Phi^{-1}(t), \tag{10}$$

where $\Phi_2(\cdot,\cdot;\rho)$ is the bivariate standard normal CDF. The Gaussian copula has $\lambda_{\mathrm{U}} = 0$ for every $\rho_S < 1$, so $\rho_{\mathrm{ind}}(t) \to 0$ and $\mathrm{DEFF}(t) \to 1$ as $t \to 1$: clustering becomes irrelevant as $\alpha \to 0$.

Corollary 2 (Tail-dependent copulas)**.** *If the within-cluster copula is Student-t with equicorrelation ρ_S and ν degrees of freedom, then $\lambda_{\mathrm{U}} = 2\,t_{\nu+1}\big(-\sqrt{(\nu+1)(1-\rho_S)/(1+\rho_S)}\big) > 0$ (Embrechts et al., 2002; Nelsen, 2006), so $\mathrm{DEFF}(t) \to 1 + (\bar{m}_a - 1)\lambda_{\mathrm{U}} > 1$ and no choice of α removes the penalty.*

Corollary 3 (Endpoints)**.** *If the $U_{j,r}$ are i.i.d., then $\rho_{\mathrm{ind}} \equiv 0$, $\mathrm{DEFF} \equiv 1$ and (5) gives $C \sim \mathrm{Beta}(k, n+1-k)$, which is Vovk (2013). If within-cluster dependence is comonotone ($U_{j,r} = V_j$ for all r), then $\lambda_{\mathrm{U}} = 1$, $\rho_{\mathrm{ind}} \equiv 1$, $\mathrm{DEFF} = \bar{m}_a$, and $N(t)$ has effective sample size b, which is the regime of Ramos et al. (2026).*

3.1 A FLOOR ON THE NUMBER OF CLUSTERS

Proposition 2 (Cluster floor for coverage)**.** *Suppose within-cluster dependence is comonotone. Then no procedure that thresholds at an observed calibration score attains $\mathbb{P}(C \ge 1-\alpha) \ge 1-\delta$ unless*

$$b \;\ge\; \frac{\log \delta}{\log(1-\alpha)}. \tag{11}$$

Proof. Under comonotonicity the n transforms take only b distinct values $V_1, \ldots, V_b \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$, each repeated m_j times. A threshold at an observed score therefore yields coverage $C = V_{(i)}$ for some $i \le b$, and the largest attainable value is $\max_j V_j$, with $\mathbb{P}(\max_j V_j < 1-\alpha) = (1-\alpha)^b$. Requiring $(1-\alpha)^b \le \delta$ gives (11). □

Corollary 4 (The same floor for risk control)**.** *Under comonotone within-cluster dependence, no procedure guarantees conformal risk control (Angelopoulos et al., 2024) at level (α, δ) for every loss distribution on $[0,1]$ unless b satisfies* (11)*.*

Proof. Take cluster-level losses $\mathrm{Bernoulli}(p)$ and compare $H_0 : p = 0$ with $H_1 : p = 1 - \delta^{1/b}$. Under H_1 all b clusters are loss-free with probability exactly δ, so the most favourable evidence arises under H_1 with probability δ and no procedure separates the hypotheses on that event. Validity therefore forces $1 - \delta^{1/b} \le \alpha$, which is (11). □

The two floors coincide, although one follows from an order statistic and the other from a zero-event argument (Clopper & Pearson, 1934).

At $\alpha = \delta = 0.1$ the floor is $b \ge 21.85$, that is, 22 clusters; at $\alpha = 0.05$ it is 45. Monte Carlo reproduces the transition, with $\mathbb{P}(C < 0.9) = 0.109$ at $b = 21$ and 0.098 at $b = 22$, matching $(1-\alpha)^b$ to four decimals. This floor also applies to federated calibration. Shahid (2026) report that pooled conformal risk control exceeds its budget at 40% of institutions across 20 hospitals and propose a site-conditional correction. At $b = 20$, Corollary 4 rules out a *distribution-free* guarantee at $(\alpha, \delta) = (0.1, 0.1)$. A site-conditional correction must therefore use additional structure on within-site dependence. The design effect in §5 makes that structure explicit and estimable.

4 THE DESIGN EFFECT IS NOT A SCALAR

Existing clustered-calibration heuristics summarise dependence with a single number, either an intraclass correlation or an interpolation weight between pooled and per-site thresholds (Shahid, 2026). Theorem 2 already shows why a single number can be insufficient: ρ_{ind} varies with the threshold t. The next construction quantifies how large the resulting discrepancy can be.

Theorem 3 (Design-effect trap)**.** *Fix $q \in (0,1)$. Let $B_j \sim \mathrm{Bernoulli}(1-q)$ be a cluster-level indicator, $\varepsilon_{j,r} \sim U(0,1)$ i.i.d., $W_j \sim U(0,1)$, and set $U_{j,r} = q\varepsilon_{j,r}$ if $B_j = 0$ and $U_{j,r} = q + (1-q)W_j$ if $B_j = 1$. Then each $U_{j,r}$ is exactly $U(0,1)$, and*

$$\rho_{\mathrm{ind}}(t) = \frac{t(1-q)}{q(1-t)} \quad (t \le q), \qquad \rho_{\mathrm{ind}}(t) = 1 \quad (t > q). \tag{12}$$

Proof. For $t \le q$, $\mathbb{P}(U \le t) = \mathbb{P}(B = 0)\mathbb{P}(q\varepsilon \le t) = q(t/q) = t$, and for $t > q$, $\mathbb{P}(U \le t) = q + (1-q)\frac{t-q}{1-q} = t$. For $t \le q$ the event $\{U \le t\}$ requires the shared $B_j = 0$ and then independent ε, so $\mathbb{P}(U \le t, U' \le t) = q(t/q)^2 = t^2/q$, which gives the stated ρ_{ind}. For $t > q$ the event $\{U > t\}$ requires $B_j = 1$ and $W_j > (t-q)/(1-q)$, both at the cluster level, so $\mathbb{P}(U > t, U' > t) = (1-q)\frac{1-t}{1-q} = 1 - t = \mathbb{P}(U > t)$ and the conditional probability is 1. □

The design effect is thus of order one below q and equal to the cluster size above it. With $q = 0.9$ the empirical ρ_{ind} matches the closed form to three decimals, and the marginals pass a Kolmogorov–Smirnov test ($p = 0.53$). A procedure that measures dependence at the median and applies it at the operating threshold under-corrects by $\mathrm{DEFF(tail)}/\mathrm{DEFF(median)}$, which equals 5.0, 6.4 and 7.8 for $\bar{m}_a = 10$, 20 and 50. The estimator in §5 accordingly evaluates ρ_{ind} at the calibration threshold.

5 A THRESHOLD-LEVEL CORRECTION

The identity in Theorem 1 yields a practical correction once ρ_{ind} is estimated from the calibration data.

Estimator. Given calibration scores grouped by cluster and an absolute threshold τ, let r_j be the fraction of cluster j at or below τ, $N = \sum_j m_j$ and $\bar{r} = \sum_j m_j r_j / N$. We use the one-way analysis-of-variance estimator for binary data (Fleiss & Cuzick, 1979),

$$\widehat{\rho}_{\mathrm{ind}}(\tau) = \frac{\mathrm{MSB} - \mathrm{MSW}}{\mathrm{MSB} + (m_0 - 1)\,\mathrm{MSW}}, \qquad m_0 = \frac{N - \sum_j m_j^2/N}{b-1}, \tag{13}$$

with $\mathrm{MSB} = \sum_j m_j (r_j - \bar{r})^2/(b-1)$ and $\mathrm{MSW} = \sum_j m_j r_j (1-r_j)/(N-b)$. Its size-adjusted denominator keeps it well defined under the very unequal cluster sizes of multi-site data, where the naive pairwise-concordance estimator can leave $[-1, 1]$.

Upper confidence bound. Because b is typically a few dozen, we replace the point estimate by an upper bound $\widehat{\rho}^{+}$, the $(1-\delta_\rho)$ quantile of the cluster bootstrap, which resamples whole clusters and so respects (A2)–(A3).

Corrected index. We seek k with $\mathbb{P}(C < 1-\alpha) \le \delta$, which by (5) is $\mathbb{P}(N(1-\alpha) \ge k) \le \delta$. Since N is a sum of b independent cluster counts, a normal approximation with the exact moments of Theorem 1 gives

$$k^\star = \left\lceil nt + z_{1-\delta}\sqrt{n\,t(1-t)\left[1 + (\bar{m}_a - 1)\widehat{\rho}^{+}\right]}\,\right\rceil + 1, \qquad t = 1-\alpha. \tag{14}$$

Setting $\widehat{\rho}^{+} = 0$ recovers the i.i.d. correction. With a union bound over the bootstrap event the nominal level is $\delta + \delta_\rho$, and we take $\delta_\rho = \delta$.

Table 1: Monte Carlo variance of $N(t)$ against the design-effect identity ($b = 400$, $m = 10$, Gaussian copula, $\rho_S = 0.5$). The indicator correlation reproduces the variance; the score correlation overstates it by 37 to 91%, increasingly in the tail.

t	MC Var$[N]$	identity, ρ_{ind}	ratio	identity, ρ_S	ratio
0.50	4013.2	4000.0	1.003	5500.0	0.730
0.80	2360.0	2337.4	1.010	3520.0	0.670
0.90	1159.0	1166.5	0.994	1980.0	0.585
0.95	548.2	538.8	1.017	1045.0	0.525

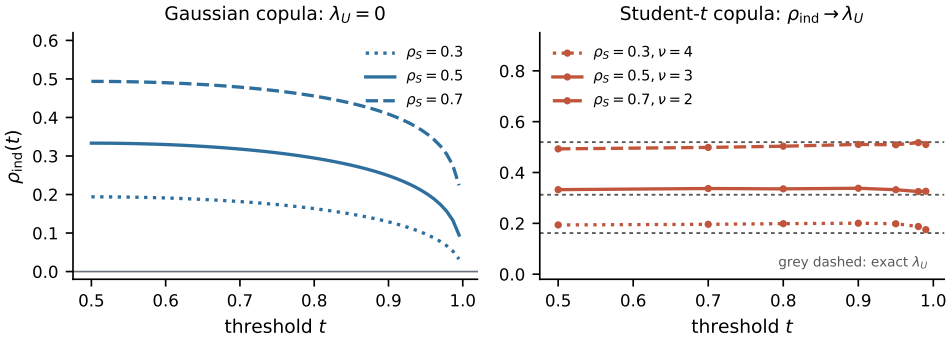

Figure 1: Indicator correlation against the threshold (Theorem 2). **Left:** under a Gaussian within-cluster copula the indicator correlation decays toward 0 as $t \to 1$ for every $\rho_S < 1$, so the design effect at the conformal threshold vanishes as $\alpha \to 0$. **Right:** under Student-t copulas it converges to the upper tail-dependence coefficient λ_{U} (grey dashed lines) and the penalty persists. The panels use the same score correlations.

6 Experiments

6.1 Simulations

Coverage variance. With $b = 24$ clusters of size 20, a Gaussian within-cluster copula at $\rho_S = 0.5$ and $\alpha = 0.1$, mean coverage is nearly unchanged (0.895 clustered against 0.900 i.i.d.), while the standard deviation of training-conditional coverage rises from 0.014 to 0.033, a roughly sixfold variance inflation, and the first percentile of realised coverage falls from 0.866 to 0.803 (Appendix A). The mean therefore masks a substantial increase in dispersion.

Exact variance and the score-correlation approximation. Table 1 compares the Monte Carlo variance of $N(t)$ with Theorem 1 evaluated at $\rho_{\text{ind}}(t)$ and at the score correlation ρ_S. The tail limit of Theorem 2 is verified in Figure 1 and Appendix B: for three Student-t copulas $\rho_{\text{ind}}(t)$ converges to the exact λ_{U} (for instance 0.326 against 0.313), while a Gaussian copula with the same ρ_S decays toward zero.

Number of clusters. Table 2 varies b at $m = 20$ under a Gaussian copula with $\rho_S = 0.5$, for which $\rho_{\text{ind}}(0.9) = 0.249$ and DEFF $= 5.73$. The nominal i.i.d. index violates the PAC target at every b. With this cluster size, additional observations within a cluster contribute much less information than observations from new clusters. The correction using the true ρ_{ind} remains within Monte Carlo error of the target throughout. The plug-in version reaches the target at about $b = 40$ and is also tighter than subsampling at larger b (calibrated level 0.918 versus 0.938 at $b = 160$).

6.2 Multi-site neuroimaging

We use ABIDE I (CPAC pipeline, CC200 atlas) (Di Martino et al., 2014; Craddock et al., 2013): 1035 subjects from 24 sites of size 6 to 175, so that the Kish mean cluster size is 72.2 against an

Table 2: PAC violation rate $\mathbb{P}(C < 1 - \alpha)$ and calibrated quantile level (lower is tighter) as the number of clusters grows, $\alpha = \delta = 0.1$, $m = 20$. * marks a violation of the PAC target.

		nominal		subsample		ours, true ρ_{ind}		ours, plug-in	
b	n	viol.	k/n	viol.	k/b	viol.	k/n	viol.	k/n
6	120	0.320*	0.942	0.527*	1.000	0.015	1.000	0.118*	0.979
20	400	0.307*	0.922	0.128*	1.000	0.088	0.950	0.112*	0.951
40	800	0.322*	0.915	0.105*	0.975	0.073	0.935	0.092	0.936
160	3200	0.293*	0.907	0.077	0.938	0.093	0.917	0.093	0.918

Table 3: ABIDE, 400 site-level splits, target 90% coverage with $\delta = 0.1$. Width is the mean calibrated half-width in years. The oracle uses the test sites and is shown as an efficiency reference.

method	mean cov.	sd	$\mathbb{P}(C < 0.9)$	width (yr)	PAC
nominal i.i.d. index	0.910	0.067	0.378	11.60	fail
subsample (Dunn et al., 2023)	0.885	0.129	0.412	14.07	fail
threshold-level correction (ours)	0.972	0.044	**0.075**	25.61	**pass**
oracle	0.907	0.010	0.007	10.65	

arithmetic mean of 43.1. Age is regressed on the 19,900 Fisher-z connectivity edges (standardisation, 100 principal components, ridge), and the score $|y - \hat{y}|$ makes the calibrated threshold the interval half-width in years. Sites are split $50/25/25$ into training, calibration and test, so each test subject comes from an unseen site as in (A4). Over 400 splits the estimated indicator correlation at the calibration threshold averages 0.151, a mean design effect of 15.3.

Table 3 separates mean coverage from the PAC criterion. The nominal index covers 91.0% of test subjects on average, but violates the PAC requirement on 37.8% of calibration draws. Subsampling violates it on 41.2% of draws because six calibration sites leave only six retained points. The threshold-level correction is the only implementable method among those compared that meets the empirical target. Its intervals are wider than the oracle's because the correction must infer coverage from only six sites.

6.3 Single-cell calibration at the donor level

Tabula Muris Senis (The Tabula Muris Consortium, 2020) contains 356,213 cells from 44 donor mice across 23 tissues; we use the atlas's own 50-dimensional PCA embedding. Predicting donor age from a single cell makes the within-donor dependence of the scores close to comonotone, because every cell of a mouse carries the same label, which places this dataset near the $\lambda_{\text{U}} = 1$ endpoint of Corollary 3. A donor-level $50/25/25$ split leaves a median of 89,569 calibration cells from about 11 donors. The estimated design effect is about 3.0×10^3, so the effective calibration sample size is

$$n_{\text{eff}} \approx 89{,}569/2966 \approx 30, \tag{15}$$

which is on the scale of the number of mice rather than the number of cells.

With about 11 calibration donors, compared with the comonotone floor of 22, every implementable method in Table 4 exceeds δ. The ordering matches the theory: the nominal index violates the target on 48.5% of splits, subsampling on 34.5%, and the threshold-level correction on 15.5%. ABIDE has much weaker dependence: the estimated indicator correlation is about 0.15, and the same correction meets the empirical target with six calibration sites. The relevant cluster requirement therefore depends on the dependence structure rather than on a universal cluster count. In single-cell conformal prediction, where donor and batch structure is pervasive (Luecken et al., 2022), the donor is the relevant calibration unit for these guarantees.

Table 4: Tabula Muris Senis, 200 donor-level splits, target 90% coverage, $\delta = 0.1$. Each split calibrates on ~ 11 donors, below the cluster floor of Proposition 2, and no implementable method attains the PAC target. Width is the mean half-width in months.

method	mean cov.	sd	$\mathbb{P}(C < 0.9)$	width (mo)
nominal i.i.d. index	0.875	0.098	0.485	16.00
subsample (Dunn et al., 2023)	0.907	0.097	0.345	17.39
threshold-level correction (ours)	0.950	0.068	**0.155**	19.43
oracle	0.900	0.000	0.000	15.94

7 Related work

Training-conditional coverage. Vovk (2013) derives the i.i.d. Beta law and the corresponding PAC correction, Bian & Barber (2023) show that training-conditional coverage holds for split conformal but not in general for full conformal or the jackknife+ (Barber et al., 2021), and Park et al. (2020) brought the (ϵ, δ) formulation to deep learning. All assume i.i.d. calibration data, and our results reduce to theirs at $\rho_{\mathrm{ind}} = 0$.

Conformal prediction under clustering. Dunn et al. (2023) obtain marginal validity by subsampling one point per cluster; Lee et al. (2025) give marginal coverage and a second-moment guarantee under hierarchical exchangeability; Dobriban & Yu (2025) and Shen & Yu (2026) treat general group symmetries, the latter targeting covariate-conditional coverage; and Lee & Ren (2025) address selection with false discovery rate control for hierarchical data. Closest to our setting, Ramos et al. (2026) transport Vovk's law to clustered data with identical within-cluster scores, which is the $\lambda_{\mathrm{U}} = 1$ endpoint of Corollary 3. Our analysis covers the intermediate regime between these endpoints and identifies tail dependence as the quantity governing the design effect. It also provides the cluster floor and the threshold-level estimator studied above.

Risk control and non-exchangeable data. Conformal risk control (Angelopoulos et al., 2024) and its high-probability counterparts (Bates et al., 2021; Angelopoulos et al., 2025b) certify monotone losses, and Farinhas et al. (2024) relax exchangeability by weighting. Barber et al. (2023) bound the coverage gap under non-exchangeability by a total-variation term, and Barber & Panan-jady (2025) characterise split conformal under temporal β-mixing; Theorem 1 is the clustered analogue and is an identity rather than a bound. Wieczorek (2023) applies survey-design reasoning, the tradition of the Kish factor (Kish, 1965), to coverage under complex sampling.

8 Limitations

Theorem 1 requires (A2)–(A3) and does not cover nested or crossed designs such as site by batch. Assumption (A4) takes the test cluster to be a fresh draw from the same population of clusters, so systematic differences at a new site, for which covariate-shift corrections (Tibshirani et al., 2019; Park et al., 2022) are the usual tool, fall outside the analysis. The corrected index (14) uses a normal approximation and a bootstrap upper bound whose justification is asymptotic in the number of clusters. In our simulations, the plug-in procedure reaches its target at about 40 clusters. Confidence sequences built by betting (Waudby-Smith & Ramdas, 2024) offer a route to a finite-sample bound on ρ_{ind} from b clusters. Proposition 2 is stated at the comonotone extreme, so for weaker dependence the floor is necessary but not tight.

Reproducibility statement

Seeded scripts in the supplementary material reproduce all reported results: `theory_check.py` and `clustered_conformal.py` for the theory checks, `impossibility.py` for the cluster floor, `deff_trap.py` for Theorem 3, `exp_sweep_table.py` for Table 2, `exp_abide_full.py` for Table 3, `exp_tms.py` for Table 4, `exact_correction.py` for Appendix C, and `make_figs.py` for the figures. `run_all.sh` reproduces the theory, simulations

and figures without external data, and `requirements.txt` pins library versions. All computations run on CPU. ABIDE I derivatives are distributed by the Preprocessed Connectomes Project and Tabula Muris Senis through CELLxGENE; both are public and the supplied loaders rebuild our inputs from the public files.

ETHICS STATEMENT

This work uses two public, de-identified research datasets under their published terms and collects no new data. The study concerns the reliability of uncertainty statements in clinical and scientific deployment, where nominal average coverage can differ substantially from site-level behavior. Age prediction from imaging or expression is used only as a methodological test case and is not proposed as a clinical tool.

AI USE STATEMENT

GPT was used only for writing and language polishing.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. Cambridge University Press, 2025a. arXiv:2411.11824.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 2025b. arXiv:2110.01052.

Rina Foygel Barber and Ashwin Pananjady. Why is split conformal prediction effective despite temporal dependence? *arXiv preprint arXiv:2510.02471*, 2025.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023. arXiv:2202.13415.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023. doi: 10.1214/23-EJS2145.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Cameron Craddock et al. The Neuro Bureau Preprocessing Initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 2013.

Adriana Di Martino et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014. doi: 10.1038/mp.2013.78.

Edgar Dobriban and Mengxin Yu. SymmPI: Predictive inference for data with group symmetries. *Journal of the Royal Statistical Society Series B*, 87(5), 2025. arXiv:2312.16160.

Allan Donner and Neil Klar. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London, 2000. ISBN 0-340-69153-0.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023. arXiv:1809.07441.

Paul Embrechts, Alexander J. McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. In M. A. H. Dempster (ed.), *Risk Management: Value at Risk and Beyond*, pp. 176–223. Cambridge University Press, 2002. doi: 10.1017/CBO9780511615337.008.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Joseph L. Fleiss and Jack Cuzick. The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement*, 3(4):537–542, 1979. doi: 10.1177/014662167900300410.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008.

Harry Joe. *Dependence Modeling with Copulas*, volume 134 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 2014. doi: 10.1201/b17116.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Yonghoon Lee and Zhimei Ren. Selection from hierarchical data with conformal e-values. *arXiv preprint arXiv:2501.02514*, 2025.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data. *ACM/IMS Journal of Data Science*, 2025. arXiv:2306.06342.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Malte D. Luecken et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022.

Scott Marek et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022. doi: 10.1038/s41586-022-04492-9.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York, 2nd edition, 2006.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, pp. 345–356, 2002.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. PAC confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations (ICLR)*, 2020.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. In *International Conference on Learning Representations (ICLR)*, 2022.

Thiago R. Ramos, Helton Graziadei, and Luben M. C. Cabezas. Conformal prediction via transported beta laws. *arXiv preprint arXiv:2605.19024*, 2026.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15, 2024.

Nafis Fuad Shahid. When average calibration fails: site-conditional federated conformal risk control. *arXiv preprint arXiv:2606.20115*, 2026.

Jiwei Shen and Mengxin Yu. Conditional predictive inference for general structured data with group symmetries. *arXiv preprint arXiv:2605.17934*, 2026.

The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817):590–595, 2020. doi: 10.1038/s41586-020-2496-1.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2–3): 349–376, 2013. doi: 10.1007/s10994-013-5355-6.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1), 2024.

Jerzy Wieczorek. Design-based conformal prediction. *arXiv preprint arXiv:2303.01422*, 2023.

A Supporting figures

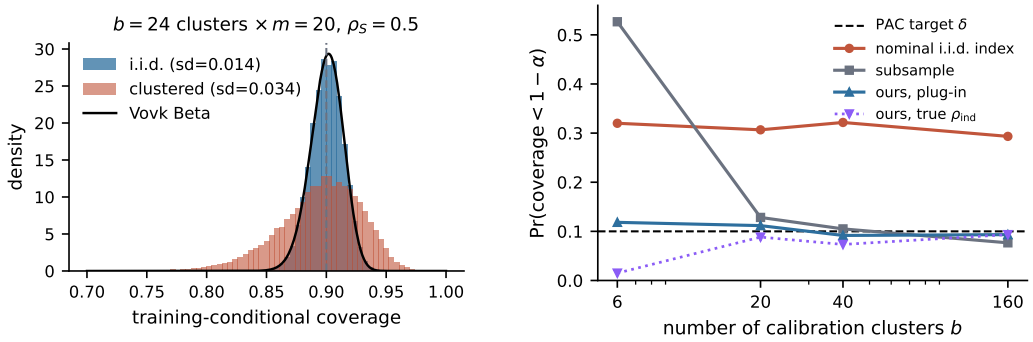

Figure 2: **Left:** training-conditional coverage under clustering is much more dispersed than Vovk's i.i.d. Beta law despite a similar mean. **Right:** PAC violation rate versus the number of calibration clusters. The nominal index exceeds the target at every b; the correction using the true ρ_{ind} stays near the target; the plug-in version reaches it near $b = 40$; and subsampling is unstable at small b because it retains only b points.

B Numerical verification of Theorem 2

Table 5: $\rho_{\mathrm{ind}}(t)$ for Student-t within-cluster copulas converges to the upper tail-dependence coefficient λ_{U}, while a Gaussian copula with the same score correlation decays toward 0 (2×10^5 clusters of size 4).

ρ_S	ν	λ_{U} (exact)	$\rho_{\mathrm{ind}}(.95)$	$\rho_{\mathrm{ind}}(.99)$	$\rho_{\mathrm{ind}}(.995)$	Gaussian $\rho_{\mathrm{ind}}(.99)$
0.5	3	0.3125	0.3331	0.3250	0.3257	0.1206
0.3	4	0.1618	0.2016	0.1845	0.1789	0.0461
0.7	2	0.5195	0.5137	0.5251	0.5133	0.2594

C AN EXACT PARAMETRIC ALTERNATIVE

A parametric alternative to the normal approximation in (14) models each cluster count as Beta-Binomial with correlation ρ_{ind} and then convolves the cluster-level count distributions. Supplied with the true ρ_{ind}, the exact route changes the calibrated quantile level only slightly (0.983 against 1.000 at $b = 6$, and 0.933 against 0.935 at $b = 40$), and its violation rate sits at the nominal level (0.107 at $b = 40$ and 0.105 at $b = 80$) rather than below it. The variance identity of Theorem 1 holds for any within-cluster exchangeable dependence, whereas the Beta-Binomial shape is an additional assumption; a correction built on the exact variance is therefore the more robust choice. The comparison indicates that, at small b, uncertainty in ρ_{ind} contributes more to the width of the corrected interval than the Gaussian approximation does.

D TEST POINT FROM A CALIBRATION CLUSTER

When the test point belongs to a cluster present in the calibration data, (A4) fails and coverage is no longer $U_{(k)}$. It becomes $\mathbb{E}[\,\mathbf{1}\{U_{\mathrm{test}} \le U_{(k)}\} \mid \mathcal{D}_{\mathrm{cal}}, j_{\mathrm{test}}]$, which depends on the within-cluster copula through the conditional law of a new draw given the observed members of its cluster. Part of the cluster effect is then shared by the test point, which attenuates the design effect, so the deployment case (A4) is the more demanding one.