BEYOND ACCURACY: CONFORMAL RISK CONTROL FOR MEDICAL TIME-SERIES CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical time-series classifiers are typically evaluated by predictive accuracy, yet accuracy alone provides no finite-sample control of clinically relevant deployment risks. We study conformal risk control (CRC) for medical time-series classification and show that its direct application is complicated by the hierarchical structure of medical time series: medical time series have a patient–window hierarchy in which each patient contributes many correlated windows. Treating windows as exchangeable calibration units creates a *nominal sample-size illusion* and may control a window-weighted risk rather than the risk for a newly sampled patient. Patient-level calibration corrects this mismatch but substantially reduces the effective calibration size, yielding a small-sample regime in which post-hoc CRC can become conservative. Separately, end-to-end conformal risk training requires careful model selection to preserve the target risk constraint. We develop classification-specific CRC for binary false-negative-rate control and multiclass coverage, together with a subject-aware formulation that calibrates directly over patients. We further extend the framework to subject-level conditional value-at-risk (CVaR) to control tail risk across patients. Across 12 architectures and three EEG/ECG datasets, experimental results reveal a validity–efficiency–stability trade-off: post-hoc CRC provides finite-sample expected-risk control under exchangeability, whereas end-to-end training can improve efficiency on binary tasks but is less efficient than closed-form fine-tuning on the multiclass ADFTD task once model selection correctly enforces the coverage constraint. Reliable deployment risk control therefore requires careful choice of both the calibration unit and the model-selection criterion used during optimization.

1 INTRODUCTION

Medical time-series classification is increasingly used for computer-assisted diagnosis from physiological signals such as electroencephalography (EEG) and electrocardiography (ECG). Recent work has substantially improved predictive performance through multi-scale temporal modeling, cross-channel interaction, self-supervised representation learning, medical-domain architectures, and foundation-model adaptation (Wang et al., 2024; 2023; Li et al., 2025; Yu et al., 2026). These models are typically trained and evaluated using empirical metrics such as accuracy, F1, recall, AUROC, and AUPRC. Such metrics characterize predictive performance on a particular test set, but do not provide finite-sample control of clinically relevant deployment risks. For example, observing a low false-negative rate (FNR) does not imply that the expected FNR of a deployed decision rule is guaranteed to remain below a prescribed tolerance. Moreover, a classifier with strong predictive performance may still operate at a threshold that is poorly aligned with a clinically specified risk constraint.

Conformal prediction (Vovk et al., 2005) and conformal risk control (CRC) (Angelopoulos et al., 2024) provide distribution-free tools for calibrating predictive decisions using held-out data. More recently, Conformal Risk Training (CRT) (Yeh et al., 2025) integrates the risk-controlling parameter into model optimization, enabling predictive utility to be optimized subject to a prescribed risk constraint. These developments provide a natural foundation for statistically controlled medical diagnosis, but medical time-series classification introduces three coupled challenges.

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.

Challenge I: From predictive performance to controlled deployment risk. Medical time-series classifiers are generally optimized for empirical predictive performance rather than explicit deployment-risk constraints. Reporting an FNR on a test set is therefore fundamentally different from constructing a decision rule whose expected FNR is controlled at a user-specified level α. Moreover, existing end-to-end conformal training formulations do not directly specify how risk calibration should be specialized to window-level binary and multiclass classification. The first challenge is thus to formulate classification-specific risk control while retaining finite-sample validity and differentiability for end-to-end optimization.

Challenge II: What is the appropriate conformal unit? Medical time-series data are hierarchical: each patient contributes multiple correlated windows, potentially originating from multiple recordings. This structure is already recognized in predictive evaluation, where subject-independent splits are commonly used to prevent patient-level leakage (Wang et al., 2024; Yu et al., 2026). For conformal calibration, however, the hierarchy determines the statistical unit to which the guarantee applies. Treating windows as exchangeable calibration observations can make a dataset with thousands of windows appear statistically large even when those windows arise from only a few dozen patients. We refer to this discrepancy as a *nominal sample-size illusion*. More importantly, the choice of calibration unit changes the estimand itself: window-level calibration weights patients according to the number of windows they contribute, whereas deployment often concerns the risk for a newly sampled patient. Consequently, control of window-weighted risk need not imply control of patient-level deployment risk.

Challenge III: Risk control with a small effective sample size. Treating patients rather than windows as the calibration units resolves this mismatch, but sharply reduces the effective calibration size. Medical time-series datasets with thousands of windows may contain only tens of exchangeable patients, placing subject-level conformal calibration in a small-sample regime obscured by the nominal window count. In this regime, finite-sample corrections become non-negligible and the set of admissible risk-controlling thresholds can be coarse, increasing conservativeness.

Separately, end-to-end CRT repeatedly estimates risk-controlling thresholds from window-level pseudo-calibration subsets during training. Its optimization and model selection can therefore be sensitive to surrogate temperature, architecture, and training seed. We study these effects separately from the small-sample behavior of subject-level calibration.

These challenges form a natural progression: deployment risk first requires an appropriate classification-specific control objective; the patient hierarchy then determines the relevant exchangeable unit; and correcting that unit exposes a small-effective-sample regime in which calibration efficiency and optimization stability become central. The same hierarchy also motivates moving beyond mean patient risk, since acceptable average performance may coexist with poor outcomes concentrated among a subset of patients.

We address these challenges with a unified framework for conformal risk control in medical time-series classification. We specialize CRC to binary FNR control and multiclass coverage, and develop differentiable calibration rules suitable for end-to-end risk training. We then introduce a *subject-aware* formulation that aggregates window-level losses within each patient and performs calibration across patients, aligning the controlled risk with deployment to a newly sampled patient. We systematically compare post-hoc and end-to-end risk control across these datasets, and separately study how subject-aware calibration behaves when the number of exchangeable patients is small.

Our contributions are:

- We develop classification-specific conformal risk control for binary FNR and multiclass coverage, together with differentiable threshold calibration for end-to-end conformal risk training.
- We introduce *subject-aware conformal risk control*, which treats the patient as the exchangeable calibration unit and averages loss across each patient's windows. This formulation exposes the *nominal sample-size illusion* of window-level calibration and directly targets risk for a newly sampled patient.

  We characterize the resulting validity–efficiency–stability trade-off across 12 published time-series architectures and three EEG/ECG datasets. Post-hoc CRC provides finite-

sample expected-risk control under exchangeability. End-to-end optimization can improve efficiency on the binary tasks, while on multiclass ADFTD it maintains the target coverage after correct model selection but does not outperform matched CE fine-tuning in prediction-set efficiency.

- We extend subject-aware CRC to *subject-level CVaR control* using the Rockafellar–Uryasev variational representation, moving from control of average patient risk toward protection of the worst-performing fraction of patients.

2 Related Work

Medical time-series classification. Medical time-series classification has advanced through multi-scale temporal modeling, cross-channel interaction, graph-based learning, self-supervised representation learning, and foundation-model adaptation. Representative methods include Medformer (Wang et al., 2024), TeCh/CoTAR (Yu et al., 2026), MedGNN (Fan et al., 2025), and COMET (Wang et al., 2023), with recent work further exploring knowledge augmentation, multimodal learning, and cross-dataset transfer (Yuan et al., 2025; Ye et al., 2026b; Huang et al., 2024; Ye et al., 2026a). These methods primarily target predictive performance. However, strong predictive performance alone does not ensure satisfaction of a prescribed deployment-risk constraint. Moreover, standard accuracy-based evaluation does not explicitly characterize the trade-off between risk control and predictive efficiency. In contrast, we study finite-sample control of clinically specified deployment risks for existing classifiers.

Additional related work on conformal prediction and risk control and on hierarchical conformal inference and patient-level risk is provided in Appendix A.

3 Preliminaries

Problem setup. Let f_θ denote a medical time-series classifier. For binary classification, $p_\theta(x) \in [0,1]$ is the positive-class probability; for K-class classification, $p_\theta(x) \in \Delta^{K-1}$ is the class-probability vector. The data contain S subjects, with subject i contributing K_i windows $\{(x_{ij}, y_{ij})\}_{j=1}^{K_i}$ and $N = \sum_{i=1}^{S} K_i$ windows in total. Any intermediate recording or trial structure is absorbed into the subject grouping and is not treated as a separate exchangeability level in our calibration analysis. A decision rule is indexed by a threshold $\lambda \in \Lambda$, and $L_i(\lambda)$ denotes the loss of calibration unit i. Given n exchangeable calibration units, our goal is to choose $\hat{\lambda}$ such that

$$\mathbb{E}\Big[L_{n+1}(\hat{\lambda})\Big] \leq \alpha \tag{1}$$

for a new exchangeable unit.

Conformal risk control and training. Conformal risk control (CRC) provides finite-sample, distribution-free control of bounded monotone losses under exchangeability (Angelopoulos et al., 2024). For the $[0,1]$-valued, nondecreasing losses considered here, using the worst-case bound $B = 1$, CRC selects

$$\hat{\lambda} = \max\left\{\lambda_{\min}, \sup\left\{\lambda : \frac{1 + \sum_{i=1}^{n} L_i(\lambda)}{n+1} \leq \alpha\right\}\right\}, \tag{2}$$

which guarantees Eq. 1 under the standard CRC assumptions. The finite-sample correction scales as $1/(n+1)$, making the number of exchangeable calibration units critical when n is small.

Conformal Risk Training (CRT) (Yeh et al., 2025) further treats $\hat{\lambda} = \hat{\lambda}(\theta)$ as part of the training computation graph. Each minibatch is partitioned into pseudo-calibration and pseudo-prediction subsets: the former determines $\hat{\lambda}(\theta)$, while the latter evaluates a differentiable utility objective through which gradients are propagated. After training, the deployment threshold is obtained from a fresh held-out calibration set.

Conditional value-at-risk. To characterize tail rather than mean risk, we use conditional value-at-risk (CVaR). For a random loss L and tail fraction $\beta \in (0, 1]$,

$$\mathrm{CVaR}_\beta(L) = \min_{t \in \mathbb{R}} \left[t + \frac{1}{\beta} \mathbb{E}\big[(L - t)_+\big] \right], \tag{3}$$

where $(u)_+ = \max(u, 0)$ (Mansini et al., 2007). Smaller β places greater emphasis on adverse outcomes, while $\beta = 1$ recovers the mean. Section 4.4 applies this formulation to subject-level diagnostic loss.

4 METHODOLOGY

We instantiate conformal risk control for medical time-series classification along three dimensions: classification-specific risk, patient-level calibration, and patient-level tail risk. We first derive calibration rules for binary FNR and multiclass coverage, then differentiate through the calibrated threshold for end-to-end training, and finally lift calibration from windows to subjects and from mean risk to CVaR.

4.1 CLASSIFICATION-SPECIFIC CONFORMAL RISK CONTROL

Binary FNR control. For binary disease detection, let $p_\theta(x) \in [0, 1]$ denote the predicted positive-class probability. Given threshold $\lambda \in [0, 1]$,

$$\hat{y}_\lambda(x) = \mathbf{1}[p_\theta(x) \geq \lambda]\,. \tag{4}$$

For a disease-positive example, the false-negative loss is

$$\ell^{\mathrm{FN}}(x; \lambda) = \mathbf{1}[p_\theta(x) < \lambda]\,, \tag{5}$$

which is bounded in $[0, 1]$ and nondecreasing in λ. We therefore seek the largest threshold satisfying the FNR constraint; among feasible thresholds, this favors fewer false-positive predictions.

Multiclass coverage control. For K classes, we define

$$C_\lambda(x) = \{k \in [K] : p_\theta(x)_k \geq \lambda\}\,, \tag{6}$$

with miscoverage loss

$$\ell^{\mathrm{cov}}(x, y; \lambda) = \mathbf{1}[y \notin C_\lambda(x)] = \mathbf{1}[p_\theta(x)_y < \lambda]\,. \tag{7}$$

Controlling its expectation at level α yields coverage at least $1 - \alpha$. As in the binary case, increasing λ improves efficiency by shrinking prediction sets, but increases risk.

Closed-form calibration. Both losses take the form

$$\ell_i(\lambda) = \mathbf{1}[u_i < \lambda]\,, \qquad u_i = \begin{cases} p_\theta(x_i), & \text{binary FNR,} \\ p_\theta(x_i)_{y_i}, & \text{multiclass coverage,} \end{cases} \tag{8}$$

where binary FNR calibration uses disease-positive examples. Let $u_{(1)} \leq \cdots \leq u_{(n)}$ denote the ordered calibration values and define

$$m_\alpha = \lfloor (n + 1)\alpha \rfloor\,. \tag{9}$$

For $\lambda_{\min} = 0$, applying the CRC rule in Eq. 2 gives

$$\hat{\lambda}(\theta) = \begin{cases} 0, & m_\alpha = 0, \\ u_{(m_\alpha)}(\theta), & m_\alpha \geq 1. \end{cases} \tag{10}$$

Thus, for the classification losses in Eq. 8, calibration reduces to a finite-sample-corrected order statistic and requires no numerical search.

Differentiation through calibration. For $m_\alpha \geq 1$, the calibrated threshold equals one selected probability,

$$\hat{\lambda}(\theta) = u_{i^\star}(\theta), \tag{11}$$

where $i^\star$ is the corresponding order-statistic index. Away from ties, $i^\star$ is locally constant almost everywhere, yielding

$$\frac{d\hat{\lambda}}{d\theta} = \frac{du_{i^\star}(\theta)}{d\theta}. \tag{12}$$

Standard sorting and indexing operations therefore retain the almost-everywhere gradient of the selected probability, without requiring a custom backward procedure (Yeh et al., 2025).

4.2 End-to-End Conformal Risk Training

Post-hoc CRC calibrates only the decision threshold and does not adapt the underlying classifier. Following CRT (Yeh et al., 2025), we instead optimize the classifier through the risk-controlling threshold. At each training step, a minibatch $\mathcal{B}$ is randomly partitioned into pseudo-calibration and pseudo-prediction subsets, $\mathcal{B}_{\text{cal}}$ and $\mathcal{B}_{\text{pred}}$. The former determines $\hat{\lambda}(\theta)$ using Eq. 10, while the latter evaluates a differentiable efficiency objective.

For binary classification, we minimize a smooth surrogate of the false-positive rate,

$$\widetilde{\mathcal{L}}_{\text{FPR}} = \frac{1}{|\mathcal{B}^-_{\text{pred}}|} \sum_{x_i \in \mathcal{B}^-_{\text{pred}}} \sigma\left(\frac{p_\theta(x_i) - \hat{\lambda}(\theta)}{T}\right), \tag{13}$$

where $\mathcal{B}^-_{\text{pred}}$ contains negative-class examples, $\sigma(\cdot)$ is the logistic sigmoid, and $T > 0$ controls surrogate smoothness.

For multiclass prediction sets, we minimize non-true-class inclusion,

$$\widetilde{\mathcal{L}}_{\text{size}} = \frac{1}{|\mathcal{B}_{\text{pred}}|(K-1)} \sum_{(x_i,y_i) \in \mathcal{B}_{\text{pred}}} \sum_{k \neq y_i} \sigma\left(\frac{p_\theta(x_i)_k - \hat{\lambda}(\theta)}{T}\right). \tag{14}$$

Because $\hat{\lambda}(\theta)$ remains in the computation graph, gradients account for both the direct effect of θ on predictions and its indirect effect through calibration.

We compare three matched regimes. POST-HOC applies CRC to a frozen conventionally trained model; CE performs additional cross-entropy fine-tuning for the same training budget before post-hoc calibration; and E2E fine-tunes through Eqs. 13 and 14. The final deployment threshold is always recalibrated on a fresh held-out calibration set. In the E2E procedure used here, pseudo-calibration and pseudo-prediction are performed at the window level; the subject-level losses in Section 4.3 are not used during E2E optimization.

4.3 Subject-Aware Conformal Risk Control

Window-level calibration treats individual windows as exchangeable units, although multiple correlated windows arise from the same patient. Let $\ell_{ij}(\lambda) \in [0, 1]$ denote the relevant window-level loss for window j of subject i, using Eq. 5 for binary FNR or Eq. 7 for multiclass coverage.

For multiclass coverage, all subjects contribute to calibration, and we define

$$L_i(\lambda) = \frac{1}{K_i} \sum_{j=1}^{K_i} \ell_{ij}(\lambda). \tag{15}$$

For binary FNR control, only disease-positive subjects contribute to the risk. Let $\mathcal{I}_+$ denote the set of disease-positive calibration subjects, let $S_+ = |\mathcal{I}_+|$, and let K_i^+ denote the number of disease-positive windows of subject i. For $i \in \mathcal{I}_+$, we define

$$L_i^{\text{FN}}(\lambda) = \frac{1}{K_i^+} \sum_{j:y_{ij}=1} \ell_{ij}^{\text{FN}}(\lambda). \tag{16}$$

In both cases, averaging within subject absorbs arbitrary within-subject dependence into a single calibration unit; however, the number of exchangeable calibration units is S for multiclass coverage and S_+ for binary FNR control.

For notational convenience, let

$$\mathcal{I}_{\text{eff}} = \begin{cases} \{1, \dots, S\}, & \text{multiclass coverage,} \\ \mathcal{I}_+, & \text{binary FNR control,} \end{cases} \tag{17}$$

and let

$$S_{\text{eff}} = |\mathcal{I}_{\text{eff}}| = \begin{cases} S, & \text{multiclass coverage,} \\ S_+, & \text{binary FNR control.} \end{cases} \tag{18}$$

Pooling windows estimates

$$\widehat{R}_{\text{win}}(\lambda) = \frac{\sum_{i=1}^{S} K_i L_i(\lambda)}{\sum_{i=1}^{S} K_i}, \tag{19}$$

which weights patients by the number of windows they contribute. In contrast,

$$\widehat{R}_{\text{subj}}(\lambda) = \frac{1}{S_{\text{eff}}} \sum_{i \in \mathcal{I}_{\text{eff}}} L_i(\lambda), \tag{20}$$

assigns equal weight to each patient.

Applying the CRC construction in Eq. 2 directly to the subject-level losses gives

$$\hat{\lambda}_{\text{subj}} = \max\left\{ \lambda_{\min}, \sup\left\{ \lambda : \frac{1 + \sum_{i \in \mathcal{I}_{\text{eff}}} L_i(\lambda)}{S_{\text{eff}} + 1} \le \alpha \right\} \right\}. \tag{21}$$

Under subject-level exchangeability, the standard CRC result yields

$$\mathbb{E}\Big[L_{\text{new}}(\hat{\lambda}_{\text{subj}}) \Big] \le \alpha. \tag{22}$$

Accordingly, empirical evaluation of the subject-level guarantee is performed by first computing the relevant loss within each evaluation subject and then averaging these subject-specific losses with equal weight across subjects, rather than pooling all evaluation windows.

Crucially, Eq. 21 depends on the number of exchangeable subject-level calibration units S_{eff}, not the nominal window count. For multiclass coverage, $S_{\text{eff}} = S$, whereas for binary FNR control, $S_{\text{eff}} = S_+$ is the number of disease-positive calibration subjects.

For multiclass coverage, ℓ_{ij} is given by Eq. 7. For binary FNR, subject-level aggregation is performed over disease-positive subjects and their disease-positive windows.

4.4 Subject-Level Tail-Risk Control

Subject-aware CRC controls mean patient risk, which may obscure poor performance concentrated among a subset of patients. We therefore extend the construction to subject-level CVaR.

For a tail fraction $\beta \in (0, 1]$, the Rockafellar–Uryasev (Rockafellar & Uryasev, 2000) representation gives

$$\text{CVaR}_\beta(L) = \min_t \left\{ t + \frac{1}{\beta} \mathbb{E}[(L - t)_+] \right\}. \tag{23}$$

Rather than inserting an empirical top-tail estimator directly into CRC, we fix t independently of the calibration sample and define

$$\widetilde{L}_{i,t}(\lambda) = t + \frac{1}{\beta} \left(L_i(\lambda) - t \right)_+. \tag{24}$$

Since $L_i(\lambda) \in [0, 1]$,

$$\widetilde{L}_{i,t}(\lambda) \le B_t := t + \frac{1}{\beta}(1 - t)_+. \tag{25}$$

For fixed t, $\widetilde{L}_{i,t}(\lambda)$ is bounded and nondecreasing in λ. Applying CRC therefore yields

$$\hat{\lambda}_{\mathrm{CVaR}} = \max\left\{\lambda_{\min}, \sup\left\{\lambda : \frac{B_t + \sum_{i=1}^{S} \widetilde{L}_{i,t}(\lambda)}{S+1} \leq \alpha\right\}\right\}. \tag{26}$$

Define

$$g_t(\lambda) = t + \frac{1}{\beta}\mathbb{E}[(L(\lambda) - t)_+]\,. \tag{27}$$

By Eq. 23,

$$\mathrm{CVaR}_\beta\Big(L(\hat{\lambda}_{\mathrm{CVaR}})\Big) = \min_{t'} g_{t'}(\hat{\lambda}_{\mathrm{CVaR}}) \leq g_t(\hat{\lambda}_{\mathrm{CVaR}}). \tag{28}$$

Thus, controlling the transformed loss in Eq. 24 provides an upper bound on subject-level CVaR.

The auxiliary parameter t must be selected independently of the calibration sample (Yeh et al., 2025). We select t using training data only: over a prespecified grid $\mathcal{T} \subset [0, \alpha)$, we retain the value yielding the largest feasible training threshold and then hold it fixed during calibration on held-out subjects. When $\beta = 1$, choosing $t = 0$ recovers the subject-level mean-risk construction in Eq. 21.

5 Experiments

5.1 Experimental Setting

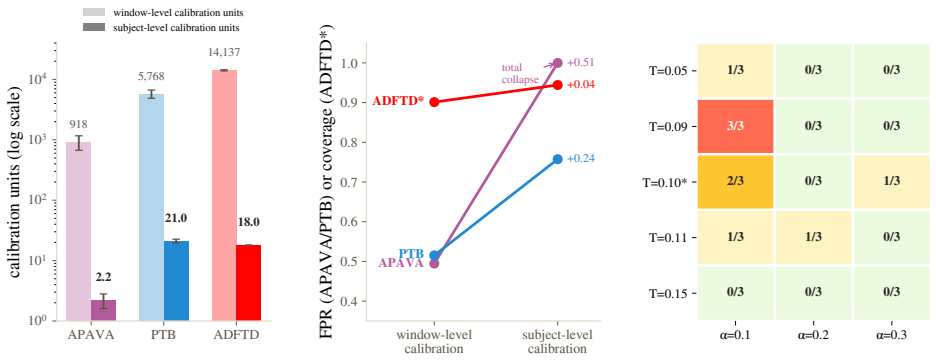
We evaluate on three subject-independent medical time-series benchmarks: **ADFTD** (three-class EEG; 88 subjects) (Miltiadous et al., 2023), **APAVA** (binary EEG; 23 subjects) (Escudero et al., 2006), and **PTB** (binary ECG; 198 subjects) (PhysioBank, 2000). All train/validation/test partitions are subject independent, so windows from the same patient never appear across partitions. Our evaluation spans 12 published backbones: `FEDformer` (Zhou et al., 2022), `TeCh` (Yu et al., 2026), `iTransformer` (Liu et al., 2024b), `Informer` (Zhou et al., 2021), `Medformer` (Wang et al., 2024), `Reformer` (Kitaev et al., 2020), `TCN` (Bai et al., 2018), `Transformer` (Vaswani et al., 2017), `PatchTST` (Nie et al., 2023), `Non-stationary Transformer` (Liu et al., 2022), `COMET` (Wang et al., 2023), and `MedGNN` (Fan et al., 2025).

We compare four regimes: uncalibrated prediction (Raw), post-hoc conformal risk control (Post-hoc) (Angelopoulos et al., 2024), matched cross-entropy fine-tuning followed by CRC (CE), and end-to-end conformal risk training (E2E) (Yeh et al., 2025), with CE controlling for gains from additional optimization. For PTB and APAVA, we control false-negative rate (FNR) and measure efficiency by false-positive rate (FPR) (Hardt et al., 2016); for ADFTD, we control miscoverage and measure efficiency by average prediction-set size. Held-out data are randomly split into calibration and evaluation sets; in subject-aware experiments, whole subjects are partitioned and losses are aggregated within subject before calibration. This keeps the held-out source fixed while varying the calibration unit.

All twelve backbones are evaluated over five independent training seeds. Implementation details, hyperparameters, calibration protocols, and statistical procedures are provided in Appendix B, with additional experimental results reported in Appendix C.

5.2 Risk Control and the Calibration Unit

Effect of risk calibration. The default operating point of an accuracy-trained classifier is not tied to a prespecified deployment-risk constraint (full results in Table 4, Appendix C.1.1). On PTB, the raw FNR ranges from 0.0210 for `FEDformer` to 0.0499 for `iTransformer`. At $\alpha = 0.05$, post-hoc CRC reduces FPR for all twelve architectures; for example, FPR decreases from 0.9363 to 0.7493 for `FEDformer` and from 0.6483 to 0.5177 for `Medformer`. Under the tighter constraint $\alpha = 0.02$, the raw FNRs of Medformer (0.0217) and `FEDformer` (0.0210) already exceed the target. These results show that predictive performance alone neither ensures satisfaction of a prescribed risk constraint nor identifies an efficient operating point under that constraint.



(a) Typical calibration units per resplit (b) Three severities, same root cause (c) ADFTD e2e: seeds collapsed / 3

Figure 1: **Calibration-unit effects and E2E stability.** (a) Typical numbers of calibration units per randomized held-out resplit. Window-level counts use the observations entering the risk calculation, whereas subject-level counts use S_{eff}: disease-positive calibration subjects for binary FNR control and all calibration subjects for multiclass miscoverage control. (b) Subject-level calibration reveals three finite-sample regimes: total collapse on APAVA, variance inflation on PTB, and an efficiency tax on ADFTD. (c) Separately, window-level E2E training on ADFTD exhibits seed-dependent instability across surrogate temperatures.

Effect of the calibration unit. The effective calibration size changes substantially when patients, rather than windows, are treated as exchangeable units. Our subject-aware experiments use randomized calibration/evaluation splits of the pooled validation/test cohorts, which contain 8 APAVA, 90 PTB, and 37 ADFTD subjects. Figure 1 summarizes the corresponding calibration regimes. For binary FNR control, the effective subject-level calibration size is smaller still because only disease-positive calibration subjects contribute, so $S_{\text{eff}} = S_+$ rather than the total number of calibration subjects.

This effect is most severe on APAVA: the small number of eligible subject-level calibration units makes the finite-sample correction too large for the tested targets, causing subject-level CRC to reduce to the trivial always-positive rule. For PTB, patient-level calibration remains informative but exhibits greater small-sample variability: at $\alpha = 0.05$, the FNR standard deviation is 0.024 under subject-mean calibration, compared with 0.049 under subsample-once calibration. For ADFTD, subject-aware calibration remains feasible but is more conservative: at $\alpha = 0.20$, average prediction-set size increases from 1.833 under window-level calibration to 2.010 under subject-level calibration, while coverage increases from 0.802 to 0.848. Additional windows improve within-patient risk estimation but do not increase the number of exchangeable patients, creating a *nominal sample-size illusion*.

5.3 Post-hoc versus End-to-End Risk Control

Table 1 summarizes efficiency at the prescribed risk targets. Because the table does not display realized risk, these efficiency differences should be interpreted together with the risk-control results below; in particular, a lower FPR or prediction-set size does not by itself imply a better operating point when the target risk is not maintained.

The effect of E2E optimization is strongly task dependent. Across the 12 architectures evaluated over five seeds, E2E yields a lower reported efficiency metric for **11/12** models on PTB and **8/12** on APAVA, but only **3/12** on ADFTD. On PTB, the largest FPR reductions are observed for `FEDformer` (-14.54pp), `Medformer` (-10.26pp), and `Non-stationary Transformer` (-8.49pp). On APAVA, the largest reductions occur for `Medformer` (-6.45pp), `Non-stationary Transformer` (-5.91pp), and `FEDformer` (-5.36pp). In contrast, E2E increases average prediction-set size for **9/12** architectures on ADFTD, with the largest increase observed for Non-stationary Transformer ($+0.852$). Only `FEDformer`, `COMET`, and `MedGNN` show smaller average set sizes under E2E, and the changes for `COMET` and `MedGNN` are negligible in magnitude.

Table 1: **Post-hoc CRC versus end-to-end CRT across architectures.** PTB and APAVA report FPR at the target FNR level $\alpha = .05$; ADFTD reports average prediction-set size at the target miscoverage level $\alpha = .20$ (target coverage 0.80). $\Delta = \text{E2E} - \text{Post}$; lower is better. All architectures report means over five independent training seeds.

Model	PTB | FPR ↓			APAVA | FPR ↓			ADFTD | Set Size ↓		
	Post	E2E	Δpp	Post	E2E	Δpp	Post	E2E	Δ
FEDformer	.747	.601	−14.54	.863	.809	−5.36	2.009	1.929	−0.079
TeCh	.389	.416	+2.76	.392	.442	+5.00	1.963	1.983	+0.019
iTransformer	.520	.491	−2.93	.528	.540	+1.16	1.817	1.847	+0.031
Informer	.438	.382	−5.60	.565	.530	−3.58	1.957	2.054	+0.097
Medformer	.504	.401	−10.26	.615	.551	−6.45	1.827	1.985	+0.158
Reformer	.450	.405	−4.53	.482	.457	−2.47	1.957	2.150	+0.194
TCN	.344	.337	−0.65	.466	.482	+1.55	1.778	2.009	+0.231
Transformer	.507	.444	−6.28	.532	.506	−2.68	1.969	2.244	+0.276
PatchTST	.470	.398	−7.21	.625	.590	−3.52	1.985	2.448	+0.463
Non-stationary Tr.	.505	.420	−8.49	.629	.570	−5.91	1.969	2.822	+0.852
COMET	.419	.406	−1.21	.634	.585	−4.89	2.166	2.154	−0.012
MedGNN	.446	.434	−1.15	.457	.495	+3.84	2.070	2.034	−0.036
Lower E2E metric	**11/12**			**8/12**			**3/12**		

Table 2: **Post-hoc, CE, and E2E on Medformer.** PTB/APAVA report FNR/FPR; ADFTD reports coverage/set size.

Dataset	α	Post	CE	E2E
PTB	.02	.020/.656	**.020/.524**	.020/.588
	.05	.051/.514	**.051/.404**	.049/.497
	.10	.101/.352	.100/.327	**.098/.217**
APAVA	.05	.046/.505	**.048/.461**	.049/.496
	.10	.094/.406	**.094/.344**	.104/.397
	.20	.191/.293	**.193/.268**	.192/.294
ADFTD	.10	.901/2.311	**.901/2.267**	.886/2.426
	.20	.802/1.833	**.798/1.738**	.785/2.059
	.30	.701/1.440	**.701/1.354**	.699/1.465

Controlling for additional fine-tuning. Unlike the cross-architecture comparison in Table 1, Table 2 is a controlled single-checkpoint comparison in which Post-hoc, CE, and E2E start from the same Medformer checkpoint. On PTB, CE is more efficient than E2E at $\alpha = .02$ and .05, whereas E2E performs best at $\alpha = .10$; on APAVA, CE yields lower FPR at all three targets. On ADFTD, after correcting the E2E model-selection procedure, all three methods attain coverage close to the requested target, but CE achieves smaller average prediction-set size than E2E at all three target levels. Thus, once model selection correctly enforces the coverage constraint, E2E maintains the target coverage on ADFTD but does not outperform matched CE fine-tuning in efficiency. The severe degradation reported in an earlier version was caused by model-selection leakage rather than by the E2E risk-training objective; see Appendix C.3.

5.4 Stability of End-to-End Risk Training

After correcting the model-selection procedure, E2E training is stable across our temperature sweeps on all three tasks. On ADFTD, the temperature- and seed-dependent collapse reported in an earlier version was caused by model-selection leakage rather than by the risk-training objective itself. After enforcing the coverage constraint during model selection and using calibration resplits disjoint from those used for selection, none of the 15 temperature–target combinations for `Medformer`, nor any of the cross-architecture combinations in Appendix C.3.4 (`COMET`, `MedGNN`), collapse or degrade.

6 Conclusion

We study conformal risk control for medical time-series classification under the patient–window hierarchy. Window-level calibration can overstate the effective sample size, while subject-level calibration better aligns with patient-level deployment at a finite-sample cost. Across 12 architectures and three EEG/ECG datasets, post-hoc CRC provides finite-sample expected-risk control under exchangeability, while E2E improves efficiency only in some regimes; on ADFTD, it maintains target coverage after correct model selection but does not outperform matched CE fine-tuning in prediction-set efficiency. Subject-level CVaR further provides tail-risk protection at the cost of increased conservativeness.

AI USE STATEMENT

Generative AI tools were used for language polishing, grammar checking, and improving clarity and readability.

REFERENCES

Anastasios Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *NeurIPS*, pp. 23047–23074, 2023.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *ICLR*, 2024.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018.

Devesh Batra, Salvatore Mercuri, and Raad Khraishi. Conformal predictions for longitudinal data. *arXiv preprint arXiv:2310.02863*, 2023.

Catherine Yu-Chi Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. Conformal tail risk control for large language model alignment. In *ICML*, 2025.

J Escudero, D Abásolo, R Hornero, P Espino, and M López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091, 2006.

Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *WWW*, pp. 5054–5064, 2025.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, pp. 3315–3323, 2016.

Nan Huang, Haishuai Wang, Zihuai He, Marinka Zitnik, and Xiang Zhang. Repurposing foundation model for generalizable medical time series classification. *CoRR*, abs/2410.03794, 2024.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: medical time series foundation model for real-world health data. In *NeurIPS*, 2025.

Yi Liu, Alexander Levis, Sharon-Lise T. Normand, and Larry Han. Multi-source conformal inference under distribution shift. In *ICML*, pp. 31344–31382, 2024a.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *NeurIPS*, pp. 9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024b.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael I. Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *ICML*, pp. 22942–22964, 2023.

Renata Mansini, Wlodzimierz Ogryczak, and Maria Grazia Speranza. Conditional value at risk and related linear programming models for portfolio optimization. *Ann. Oper. Res.*, 152(1):227–256, 2007.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6): 95, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *ECML*, pp. 345–356, 2002.

PhysioToolkit PhysioBank. Physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215 – e215, 2000.

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

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *ICLR*, 2022.

Sophia Sun and Rose Yu. Conformal prediction for time-series forecasting with change points. In *NeurIPS*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

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

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. In *NeurIPS*, pp. 55694–55717, 2023.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *NeurIPS*, pp. 36314–36341, 2024.

Jiexia Ye, Jia Li, Chang Xu, and Fugee Tsung. MedTVL: Harnessing vision and language for medical time series classification. In *KDD*, pp. 12691–12702, 2026a.

Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, and Fugee Tsung. Medspaformer: A transferable transformer with multi-granularity token sparsification for medical time series classification. In *AAAI*, pp. 27791–27799, 2026b.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *NeurIPS*, 2025.

Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica I. Avilés-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. *CoRR*, abs/2602.18473, 2026.

Xiaoyan Yuan, Wei Wang, Junxin Chen, and Xiping Hu. Reading between the channels: Knowledge-augmented medical time series classification. In *MM*, pp. 8978–8987, 2025.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *ICML*, pp. 27268–27286, 2022.

A Related Work

Conformal prediction and risk control. Conformal prediction provides distribution-free marginal coverage under exchangeability (Vovk et al., 2005; Papadopoulos et al., 2002), while conformal risk control (CRC) extends this principle to bounded monotone losses (Angelopoulos et al., 2024). ConfTr (Stutz et al., 2022) introduced differentiable conformal training for classification sets, and Conformal Risk Training (CRT) (Yeh et al., 2025) differentiates through the risk-controlling parameter to optimize utility under explicit risk constraints, including optimized certainty-equivalent risks such as CVaR. These methods typically treat the calibration observations themselves as the relevant exchangeable units. This assumption becomes nontrivial for medical time series, where many correlated windows may originate from the same patient. Our setting therefore requires distinguishing nominal window counts from the number of independent calibration units.

Hierarchical conformal inference and patient-level risk. Conformal methods for dependent data include time-series approaches for temporal dependence and distribution shift (Angelopoulos et al., 2023; Sun & Yu, 2025), hierarchical and longitudinal constructions for repeated observations (Robin et al., 2023; Batra et al., 2023), and methods for partial exchangeability or multi-source heterogeneity (Lu et al., 2023; Liu et al., 2024a). These works highlight that dependence and grouping affect the exchangeability structure underlying conformal guarantees. Our target differs: multiple windows are observed per patient, whereas deployment risk is defined for a newly sampled patient. We therefore aggregate losses within patients and apply CRC across patients. This aligns the calibration unit with the patient-level deployment target. We further extend this construction from mean patient risk to patient-level CVaR, building on classical tail-risk formulations and recent conformal tail-risk control (Mansini et al., 2007; Yeh et al., 2025; Chen et al., 2025).

B Additional Experimental Details

B.1 Datasets and Subject-Independent Evaluation

Table 3 summarizes the three medical time-series classification benchmarks. All experiments use subject-independent partitions: all windows originating from the same patient remain within a single training, validation, or test partition.

Table 3: Datasets used in our experiments.

Dataset	Modality	Task	Subjects	Windows	Classes
ADFTD	EEG	CN / FTD / AD	88	69,752	3
APAVA	EEG	AD / Healthy	23	5,967	2
PTB	ECG	MI / Normal	198	64,356	2

The original test partitions contain 1,431 windows from only 4 APAVA (Escudero et al., 2006) subjects, 9,368 windows from 61 PTB (PhysioBank, 2000) subjects, and 14,648 windows from 19 ADFTD (Miltiadous et al., 2023) subjects. The corresponding pooled validation/test cohorts contain 8, 90, and 37 subjects, respectively. These test-partition counts are reported descriptively; the conformal experiments use randomized calibration/evaluation splits of the pooled validation/test cohorts described in Appendix B.4. The large discrepancy between the nominal number of windows and the number of independent subjects motivates the subject-aware analysis in the main text.

B.2 Backbones

We evaluate 12 published time-series architectures: `FEDformer` (Zhou et al., 2022), `TeCh` (Yu et al., 2026), `iTransformer` (Liu et al., 2024b), `Informer` (Zhou et al., 2021), `Medformer` (Wang et al., 2024), `Reformer` (Kitaev et al., 2020), `TCN` (Bai et al., 2018), `Transformer` (Vaswani et al., 2017), `PatchTST` (Nie et al., 2023), `Non-stationary Transformer` (Liu et al., 2022), `COMET` (Wang et al., 2023), and `MedGNN` (Fan et al., 2025).

All twelve architectures are evaluated over five independent training seeds in the main cross-architecture comparison. Together, the evaluated backbones span frequency-domain modeling, con-

ventional and efficient Transformers, convolutional temporal modeling, medical-specific architectures, graph-based learning, hierarchical self-supervised representation learning, and foundation-model adaptation.

B.3 Compared Methods

Raw. The conventionally trained classifier is evaluated using its default decision rule without conformal calibration.

Post-hoc. The predictive model is frozen and CRC is applied to a held-out calibration set to obtain the risk-controlling threshold (Angelopoulos et al., 2024).

CE. Starting from the same conventionally trained checkpoint, the model receives additional cross-entropy fine-tuning before post-hoc CRC is applied. This baseline controls for the effect of additional optimization (Angelopoulos et al., 2024).

E2E. The model is optimized using end-to-end conformal risk training. After optimization, a fresh held-out calibration set is used to obtain the final deployment threshold.

For subject-aware experiments, we additionally compare our subject-mean construction against the hierarchical subsample-once construction of Robin et al. (2023).

B.4 Calibration and Evaluation Protocol

The original validation and test partitions are excluded from model training. For conformal evaluation, we pool these two held-out partitions and repeatedly construct randomized calibration/evaluation splits using the original validation-to-test ratio. Unless otherwise stated, calibration results are averaged over 10 random resplits.

Reporting across training seeds. For the main cross-architecture comparison, all twelve backbones are reported as means over five independent training seeds, with calibration metrics averaged over the randomized held-out resplits within each seed.

The Post-hoc/CE/E2E Medformer ablation in Table 2 and Table 9 uses a single common reference checkpoint so that the effect of the fine-tuning objective is isolated. Consequently, its numerical values need not equal the five-seed `Medformer` means reported in the cross-architecture tables.

For window-level experiments, individual windows are used as calibration units. For subject-aware experiments, entire patients are resampled, with all windows from the same patient kept together.

Subject-level evaluation follows the same unit as the calibration guarantee: the relevant metric is first computed within each evaluation subject and then averaged equally across subjects, rather than by pooling all evaluation windows.

For binary FNR control, calibration is performed only over disease-positive subjects using their disease-positive windows. The controlled FNR risk is evaluated analogously by averaging subject-specific FNRs equally across disease-positive evaluation subjects. Consequently, the effective calibration size is S_+, the number of disease-positive calibration subjects, rather than the total number of calibration subjects.

More generally, the effective calibration size in subject-aware experiments is the number of exchangeable subjects relevant to the target risk: all calibration subjects for multiclass coverage and disease-positive calibration subjects for binary FNR control.

Evaluation of subject-level CVaR. For CVaR experiments, tail-risk metrics are computed from the subject-specific evaluation losses. Let $\mathcal{I}_{\mathrm{eval}}$ denote the evaluation subjects relevant to the target risk and let $S_{\mathrm{eval}} = |\mathcal{I}_{\mathrm{eval}}|$. We report the empirical subject-level CVaR

$$\widehat{\mathrm{CVaR}}_\beta = \min_{t' \in [0,1]} \left\{ t' + \frac{1}{\beta S_{\mathrm{eval}}} \sum_{i \in \mathcal{I}_{\mathrm{eval}}} \left(L_i(\hat{\lambda}_{\mathrm{CVaR}}) - t' \right)_+ \right\}, \tag{29}$$

together with the empirical transformed risk evaluated at the training-selected, calibration-independent threshold t,

$$\widehat{g}_t = t + \frac{1}{\beta S_{\text{eval}}} \sum_{i \in \mathcal{I}_{\text{eval}}} \big(L_i(\hat{\lambda}_{\text{CVaR}}) - t\big)_+. \tag{30}$$

The former directly summarizes realized tail risk across evaluation subjects, whereas the latter is the empirical counterpart of the transformed quantity controlled by the CRC construction. For binary FNR control, these quantities are computed over disease-positive evaluation subjects using subject-specific FNR losses; for multiclass coverage, they are computed over all evaluation subjects using subject-level miscoverage losses.

The pooled-resplit protocol reduces dependence on one particular held-out subject ordering. It should not be interpreted as establishing exchangeability with an external clinical population.

B.5 Implementation Details

All experiments are run inside a Singularity container (Python 3.10.15, PyTorch 2.4.1, CUDA 12.1) on a shared SLURM-managed Linux cluster, scheduled across a heterogeneous mix of GPU types (NVIDIA T4, A10, L4, L40S, and A40) depending on availability at submission time; no result depends on a specific GPU model. Models are optimized using Adam with an initial learning rate of 10^{-4}, no weight decay, and batch size 128 (architecture-specific deviations, where used, are given in Appendix B.2). Conventional training is run for up to 30 epochs with a constant learning rate and early stopping (patience 8 epochs, no improvement in validation loss).

For the controlled CE/E2E fine-tuning comparison, both methods start from the same trained checkpoint and use the same fine-tuning budget of up to 50 epochs with early stopping (patience 10 epochs, model selection on pooled-resplit validation FPR), with identical optimizer, learning rate (10^{-4}), batch size, and data exposure. For E2E training, each minibatch is randomly partitioned in half into pseudo-calibration and pseudo-prediction subsets, i.e. pseudo-calibration fraction $\rho_{\text{cal}} = .5$.

For subject-level CVaR calibration, the auxiliary threshold t is selected using training data only from the prespecified grid

$$\mathcal{T}_{\text{CVaR}} = \left\{ \tfrac{k}{24}\,\alpha \;:\; k = 0, 1, \ldots, 23 \right\} \subset [0, \alpha),$$

i.e. 24 equally spaced candidates strictly below the target level α, and is held fixed before conformal calibration on held-out subjects.

C Additional Experimental Results

C.1 Risk Control and the Calibration Unit

C.1.1 Raw versus Calibrated Prediction

Table 4 gives the complete raw-versus-post-hoc comparison underlying Section 5.2. Post-hoc calibration reduces FPR for all twelve tested architectures at $\alpha = .05$. The gain is largest when the uncalibrated operating point is unnecessarily conservative. For example, `FEDformer` recovers 18.7pp FPR, whereas `iTransformer` changes by only 0.5pp because its raw FNR is already near the target. At the tighter target $\alpha = .02$, the raw FNRs of `Medformer` (.0217) and `FEDformer` (.0210) exceed the requested level.

Overall, calibration mainly converts excess conservativeness in the raw decision rule into lower FPR while moving FNR toward the prescribed boundary. The magnitude of the gain therefore depends strongly on how far the original operating point lies from the target risk level.

C.1.2 Window- versus Subject-Level Calibration

Table 5 compares window- and subject-level calibration, with subject-level evaluation metrics computed by equal-weight averaging over evaluation subjects, matching the estimand in Eq. 22.

Table 4: Raw versus post-hoc CRC on PTB at $\alpha = .05$.

Model	Raw FNR	Raw FPR	Post-hoc FNR/FPR	ΔFPR
FEDformer	.0210	.9363	.0493/.7493	−18.7pp
Medformer	.0217	.6483	.0495/.5177	−13.1pp
Transformer	.0331	.5931	.0487/.4870	−10.6pp
Informer	.0334	.5707	.0499/.4743	−9.6pp
PatchTST	.0300	.6049	.0485/.5181	−8.7pp
Non-stationary Tr.	.0373	.6152	.0489/.5308	−8.4pp
Reformer	.0380	.4662	.0508/.3836	−8.3pp
TCN	.0291	.3931	.0506/.3217	−7.1pp
TeCh	.0327	.5422	.0498/.4730	−6.9pp
iTransformer	.0499	.4228	.0506/.4176	−0.5pp
COMET	.0346	.3564	.0503/.2865	−7.0pp
MedGNN	.0365	.4747	.0505/.3923	−8.2pp

Table 5: Window-level versus subject-level calibration. PTB/APAVA report FNR/FPR; ADFTD reports coverage/average prediction-set size. Window metrics pool all evaluation windows, whereas Subject mean and Subsample-once first compute each subject's metric and then average equally across subjects. Subject-level entries are reported as mean ± standard deviation across resplits.

Dataset	α	Window	Subject mean	Subsample-once
APAVA	.05	.0489/.4946	.0000/1.0000	.0000/1.0000
	.10	.0995/.4006	.0000/1.0000	.0000/1.0000
	.20	.1955/.2910	.0000/1.0000	.0000/1.0000
PTB	.05	.0513/.5152	.0261±.024/.7575±.118	.0585±.049/.6360±.175
	.10	.1031/.3499	.0845±.049/.5121±.111	.1010±.063/.4521±.159
	.20	.2019/.1873	.1782±.063/.2643±.093	.1917±.095/.2743±.151
ADFTD	.10	.9012/2.3137	.9443±.042/2.6219±.233	.9616±.033/2.7487±.231
	.20	.8020/1.8325	.8475±.054/2.0100±.155	.8494±.092/2.0721±.393
	.30	.7017/1.4416	.7526±.076/1.6012±.135	.7378±.086/1.5470±.228

APAVA: finite-sample collapse. The pooled held-out cohort contains only eight subjects. For binary FNR control, however, the relevant effective calibration size is the number S_+ of disease-positive calibration subjects, rather than the total number of calibration subjects. Across our randomized resplits, S_+ ranges from 1 to 3 disease-positive subjects (mean 2.2 ± 0.6). Consequently, the CRC finite-sample correction is $1/(S_+ + 1)$, which is already large relative to the tested FNR targets. The resulting subject-level CRC therefore reduces to the trivial always-positive rule under both subject-level constructions.

PTB: variance inflation. With 90 pooled subjects, subject-aware calibration remains informative. At $\alpha = .05$, the FNR standard deviation of the subject-mean construction is $.024$, compared with $.049$ for subsample-once calibration. The corresponding FPR standard deviations are $.118$ and $.175$. Using every window from each patient therefore reduces resampling variability, although it does not uniformly minimize mean FPR.

ADFTD: efficiency tax. Subject-aware calibration remains feasible but becomes conservative. At $\alpha = .20$, window-level CRC obtains coverage $.802$ with average set size 1.833, whereas subject-mean calibration obtains coverage $.848$ with size 2.010.

The comparison with subsample-once calibration is mixed: subject averaging is more efficient at $\alpha = .10$ and $.20$, whereas subsample-once is slightly more efficient at $\alpha = .30$.

These results show that changing the calibration unit affects both the effective sample size and the operating point of the calibrated predictor. Subject-level calibration better matches patient-level deployment, but the resulting reduction in exchangeable units can manifest as collapse, higher variance, or an efficiency penalty depending on the dataset.

C.2 POST-HOC VERSUS END-TO-END RISK CONTROL

The main text reports the mean post-hoc and E2E performance of all 12 architectures in Table 1. Here we report multi-seed variability, directional consistency, and paired statistical tests for the 12-backbone five-seed cohort.

C.2.1 PTB

Table 6: Post-hoc versus E2E on PTB at $\alpha = .05$. Δ denotes E2E minus post-hoc FPR in percentage points. All models report results over five independent training seeds.

Model	Post-hoc	E2E	ΔFPR (pp)	Lower FPR	p
FEDformer	.7465	.6011	-14.54 ± 10.37	4/5	.1250
Medformer	.5036	.4010	-10.26 ± 6.34	5/5	.0625
Non-stationary Tr.	.5048	.4200	-8.49 ± 1.00	5/5	.0625
PatchTST	.4699	.3977	-7.21 ± 6.22	4/5	.1250
Transformer	.5069	.4441	-6.28 ± 8.51	4/5	.1250
Informer	.4377	.3816	-5.60 ± 5.24	4/5	.1250
Reformer	.4498	.4045	-4.53 ± 4.20	4/5	.1250
iTransformer	.5201	.4908	-2.93 ± 3.13	5/5	.0625
TCN	.3437	.3373	-0.65 ± 3.71	3/5	.8125
TeCh	.3885	.4160	$+2.76 \pm 6.34$	2/5	.4375
COMET	.4186	.4065	-1.21 ± 2.85	4/5	.4375
MedGNN	.4457	.4343	-1.15 ± 7.54	4/5	.6250

Table 6 shows that E2E yields a lower mean FPR for 11/12 architectures. The largest average reductions are observed for FEDformer, Medformer, and Non-stationary Transformer.

The pattern is broad but not uniform. Several architectures show sizeable mean reductions together with substantial seed-to-seed variation, indicating that PTB is comparatively favorable to E2E optimization while the effect remains architecture dependent.

C.2.2 APAVA

Table 7: Post-hoc versus E2E on APAVA at $\alpha = .05$. Δ denotes E2E minus post-hoc FPR in percentage points. All models report results over five independent training seeds.

Model	Post-hoc	E2E	ΔFPR (pp)	Lower FPR	p
Medformer	.6150	.5505	-6.45 ± 6.36	5/5	.0625
Non-stationary Tr.	.6290	.5700	-5.91 ± 6.18	5/5	.0625
FEDformer	.8630	.8094	-5.36 ± 4.33	4/5	.1250
Informer	.5653	.5295	-3.58 ± 3.07	5/5	.0625
PatchTST	.6252	.5900	-3.52 ± 13.75	4/5	.4375
Transformer	.5323	.5055	-2.68 ± 2.22	5/5	.0625
Reformer	.4816	.4569	-2.47 ± 2.66	4/5	.1250
iTransformer	.5284	.5400	$+1.16 \pm .86$	0/5	.0625
TCN	.4664	.4818	$+1.55 \pm 1.26$	0/5	.0625
TeCh	.3915	.4415	$+5.00 \pm 4.12$	1/5	.1250
COMET	.6338	.5849	-4.89 ± 2.98	5/5	.0625
MedGNN	.4570	.4954	$+3.84 \pm 4.77$	0/5	.0625

Table 7 shows that E2E yields a lower mean FPR for 8/12 architectures, with heterogeneous effect sizes across backbones. Compared with PTB, the reductions are generally more modest and the direction of change is less consistent across architectures. This suggests that E2E training can improve efficiency on APAVA, but the benefit remains backbone dependent rather than universal.

C.2.3 ADFTD

Table 8 shows a markedly different pattern on ADFTD. E2E yields a lower mean prediction-set size for only 3/12 architectures, while the remaining 9/12 exhibit an increase on average. Among the

Table 8: Post-hoc versus E2E on ADFTD at $\alpha = .20$ (target coverage 0.80). Δ denotes E2E minus post-hoc average prediction-set size. All models report results over five independent training seeds.

Model	Post-hoc	E2E	Δ size	Larger size	p
FEDformer	2.0088	1.9294	$-0.0794 \pm .156$	2/5	.4375
TeCh	1.9633	1.9826	$+0.0193 \pm .180$	3/5	.8125
iTransformer	1.8170	1.8474	$+0.0305 \pm .027$	5/5	.0625
Informer	1.9571	2.0543	$+0.0972 \pm .144$	3/5	.3125
Medformer	1.8269	1.9849	$+0.1579 \pm .116$	5/5	.0625
Reformer	1.9566	2.1504	$+0.1938 \pm .514$	2/5	1.0000
TCN	1.7784	2.0090	$+0.2306 \pm .541$	3/5	.8125
Transformer	1.9685	2.2440	$+0.2755 \pm .423$	5/5	.0625
PatchTST	1.9847	2.4476	$+0.4629 \pm .507$	4/5	.1250
Non-stationary Tr.	1.9694	2.8215	$+0.8521 \pm .400$	5/5	.0625
COMET	2.1662	2.1544	$-0.0119 \pm .011$	0/5	.0625
MedGNN	2.0698	2.0341	$-0.0357 \pm .093$	1/5	.4375

original backbones, iTransformer, Medformer, Transformer, and Non-stationary Transformer show larger set sizes in all five seeds and attain the smallest possible exact two-sided Wilcoxon p-value with five paired observations, 0.0625.

The contrast with the two binary datasets is pronounced: on multiclass ADFTD, E2E rarely reduces average prediction-set size and often increases it consistently across seeds. Together with the realized-risk results reported elsewhere, this cross-architecture evidence supports a task-dependent pattern rather than a general efficiency advantage of E2E training.

C.2.4 Training-Seed Robustness

Expanding the original single-seed comparison to five independent training seeds changes the qualitative win/no-win conclusion for several architectures. On PTB, Transformer and Reformer become consistent directional improvements, whereas TeCh does not. This motivates basing the main cross-architecture conclusions on the five-seed cohort rather than on one checkpoint.

More generally, the change from single-seed to multi-seed conclusions shows that optimization noise can be comparable to the method effect for some architectures. Directional consistency across seeds is therefore more informative here than the outcome of any individual checkpoint.

C.2.5 Controlling for Additional Fine-Tuning

Table 9 provides a controlled comparison of post-hoc CRC, additional CE fine-tuning followed by CRC, and E2E CRT starting from the same Medformer checkpoint.

The comparison is deliberately mixed. On PTB, CE is more efficient than E2E at $\alpha = .02$ and $.05$, whereas E2E is most efficient at $\alpha = .10$. On APAVA, CE achieves lower FPR at all three tested targets. On ADFTD, all three methods attain realized coverage close to the requested target after correcting the E2E model-selection procedure. CE nevertheless achieves smaller average prediction-set size than E2E at $\alpha = .10$ and $.20$, while their efficiencies are similar at $\alpha = .30$.

Thus, additional optimization alone does not yield a systematic efficiency advantage for E2E risk training across datasets. The CE control also separates two effects that would otherwise be confounded: some efficiency gains can indeed be reproduced by ordinary fine-tuning, while the previously observed ADFTD failure was an artifact of the flawed model-selection procedure rather than a property of the E2E risk-training objective.

An earlier version of this comparison reported severe E2E coverage degradation on ADFTD at $\alpha = .10$ and $.20$. As detailed in Section C.3, this behavior was caused by the model-selection procedure rather than by the E2E risk-training objective: checkpoint selection did not enforce the coverage constraint, and the resplits used for model selection were not fully separated from those used for final evaluation. After correcting these issues, the severe degradation disappears.

Table 9: Post-hoc CRC, CE fine-tuning, and E2E training on `Medformer`. PTB/APAVA entries are FNR/FPR; ADFTD entries are coverage/average prediction-set size. All methods start from the same reference Medformer checkpoint (single training seed), and reported values are averaged over 10 randomized calibration/evaluation resplits. This controlled comparison is separate from the five-seed cross-architecture means reported in Tables 1 and 6–8.

Dataset	α	Post-hoc	CE	E2E
PTB	.02	.0199/.6561	.0196/.5235	.0199/.5880
PTB	.05	.0505/.5143	.0513/.4037	.0494/.4973
PTB	.10	.1009/.3516	.1001/.3265	.0980/.2167
APAVA	.05	.0455/.5046	.0484/.4612	.0488/.4956
APAVA	.10	.0939/.4061	.0944/.3443	.1037/.3967
APAVA	.20	.1913/.2926	.1929/.2679	.1920/.2943
ADFTD	.10	.9005/2.3106	.9005/2.2671	.886/2.426
ADFTD	.20	.8021/1.8334	.7981/1.7383	.785/2.059
ADFTD	.30	.7007/1.4401	.7012/1.3539	.699/1.465

C.3 Stability of End-to-End Risk Training

We study sensitivity to the smooth-surrogate temperature $T \in \{.05, .09, .10, .11, .15\}$. For `Medformer`, each temperature is evaluated over three independent training seeds and three target levels, yielding 45 E2E retrainings per dataset.

C.3.1 PTB

Table 10: `Medformer` E2E temperature sensitivity on PTB. Entries are mean $\pm$ standard deviation FNR/FPR over three seeds.

T	$\alpha = .05$	$\alpha = .10$	$\alpha = .20$
.05	.0494±.0035/.4098±.0318	.1006±.0053/.2328±.0232	.2010±.0062/.0876±.0087
.09	.0500±.0032/.4210±.0133	.1000±.0049/.2456±.0128	.2009±.0059/.0974±.0076
.10	.0502±.0034/.4082±.0151	.0998±.0051/.2513±.0320	.2007±.0053/.0907±.0182
.11	.0501±.0037/.4178±.0391	.1001±.0055/.2426±.0212	.2000±.0044/.1101±.0119
.15	.0496±.0034/.4364±.0114	.0996±.0041/.2612±.0157	.2008±.0072/.0950±.0090

Table 10 shows that E2E training on PTB is comparatively insensitive to the surrogate temperature over the tested range. Across all five temperatures, realized FNR remains tightly concentrated around the prescribed target: approximately .05, .10, and .20 for the three risk levels, respectively.

Temperature primarily affects efficiency rather than realized risk. Across $T \in \{.05, .09, .10, .11, .15\}$, mean FPR ranges from .4082 to .4364 at $\alpha = .05$, from .2328 to .2612 at $\alpha = .10$, and from .0876 to .1101 at $\alpha = .20$. Moreover, the FPR changes are not monotonic in T, and the temperature giving the lowest FPR depends on the target risk level.

Thus, within the tested range, PTB exhibits a stable risk-control regime in which the surrogate temperature induces moderate efficiency variation without causing systematic loss of the target FNR constraint. Section C.3.3 reports the corresponding ADFTD sweep, which, once the model-selection procedure is corrected (see that section), is similarly stable across the tested temperature–target grid.

C.3.2 APAVA

Table 11 shows a similarly stable pattern on APAVA. Across all tested temperatures, realized FNR remains close to the prescribed target, ranging from .0474 to .0522 at $\alpha = .05$, from .0944 to .0993 at $\alpha = .10$, and from .1876 to .1971 at $\alpha = .20$.

Temperature primarily affects efficiency rather than realized risk. Across $T \in \{.05, .09, .10, .11, .15\}$, mean FPR ranges from .4632 to .4717 at $\alpha = .05$, from .3695 to .3867 at $\alpha = .10$, and from .2596 to .2926 at $\alpha = .20$. The dependence on T is again non-monotonic: the lowest FPR is obtained at $T = .11$ for $\alpha = .05$ and .20, but at $T = .10$ for $\alpha = .10$.

Table 11: `Medformer` E2E temperature sensitivity on APAVA. Entries are mean ± standard deviation FNR/FPR over three seeds.

T	$\alpha = .05$	$\alpha = .10$	$\alpha = .20$
.05	.0522±.0102/.4709±.0249	.0944±.0155/.3867±.0188	.1914±.0207/.2926±.0169
.09	.0504±.0099/.4717±.0241	.0976±.0141/.3719±.0337	.1971±.0152/.2740±.0303
.10	.0500±.0078/.4709±.0402	.0980±.0155/.3695±.0255	.1876±.0175/.2852±.0280
.11	.0474±.0111/.4632±.0232	.0993±.0094/.3850±.0224	.1953±.0166/.2596±.0199
.15	.0476±.0101/.4667±.0256	.0966±.0137/.3717±.0253	.1962±.0205/.2610±.0177

Thus, APAVA exhibits a stable binary-task regime in which surrogate temperature changes the operating efficiency without causing a systematic departure from the target FNR. Together with the PTB results in Table 10, this provides a binary-task baseline that is directly comparable to the ADFTD results in Section C.3.3.

C.3.3 ADFTD

An earlier version of this evaluation reported collapse (empirical coverage below .5) in 7 of the 15 temperature–seed combinations at $\alpha = .10$, with additional isolated collapses at $\alpha = .20$ and $\alpha = .30$. We traced this to two compounding issues in the E2E model-selection procedure (Algorithm 2 as implemented): (i) the epoch checkpoint was selected purely by minimizing average prediction-set size on a per-epoch validation resplit, with no check that realized coverage stayed near the target, so a checkpoint that shrank sets by sacrificing coverage could be selected outright; and (ii) the per-epoch validation resplits and the resplits used to compute the final reported number were drawn from the same fixed random-number stream (both seeded from 0), so up to half of the "final" resplits were the exact calibration/evaluation split already used to pick that checkpoint, violating the assumption that the model is fixed independently of the calibration data. We corrected both issues: model selection now penalizes any coverage shortfall beyond a small tolerance in addition to set size, and the final resplits are drawn from a disjoint segment of the random-number stream from the one used during selection. Table 12 reports the corrected sweep: across all 15 temperature–target combinations, not a single seed collapses or falls outside ±.02 of its target coverage. Temperature still affects efficiency: mean set size at $\alpha = .10$ ranges from 2.3813 at $T = .10$ to 2.4319 at $T = .09$, a substantially narrower range than under the uncorrected procedure.

Table 12: `Medformer` E2E temperature sensitivity on ADFTD, after correcting the model-selection leakage described in the text. "Collapsed" counts runs with empirical coverage below .5. Coverage and set size are averaged over the (here, all) non-collapsed seeds.

T	α	Collapsed	Coverage	Size
.05	.10	0/3	.9009±.0012	2.3668±.1137
	.20	0/3	.8002±.0015	1.9191±.1572
	.30	0/3	.7013±.0012	1.4446±.0195
.09	.10	0/3	.8846±.0047	2.4319±.0355
	.20	0/3	.7998±.0022	1.9631±.1318
	.30	0/3	.7007±.0008	1.4907±.0456
.10	.10	0/3	.9002±.0009	2.3813±.1062
	.20	0/3	.8006±.0018	2.0575±.2018
	.30	0/3	.7009±.0013	1.4662±.0125
.11	.10	0/3	.9012±.0010	2.4036±.0955
	.20	0/3	.8009±.0009	1.8697±.0477
	.30	0/3	.7009±.0006	1.4827±.0153
.15	.10	0/3	.9000±.0007	2.4241±.0822
	.20	0/3	.8001±.0013	1.9399±.1478
	.30	0/3	.7016±.0015	1.5094±.0644

Table 13: Cross-architecture ADFTD stability at $\alpha = .10$, after correcting the model-selection leakage described in Section C.3.3. Entries count failed seeds among three runs. C denotes collapsed (coverage < 0.5) and D denotes degraded (coverage 0.5–0.85) runs; "–" indicates no failed seed.

Architecture	$T = .05$	$T = .09$	$T = .10$	$T = .11$	$T = .15$
COMET	–	–	–	–	–
MedGNN	–	–	–	–	–
Medformer	–	–	–	–	–

C.3.4 Cross-Architecture Stability on ADFTD

An earlier version of Table 13 reported substantial architecture-dependent instability: MedGNN with 6/15 degraded runs and Medformer collapsing in 7/15 combinations. Once the model-selection leakage described in Section C.3.3 is corrected, none of the three architectures evaluated here – COMET, MedGNN, and Medformer – exhibit a single collapsed or degraded run across the full temperature–seed grid. This indicates that the previously observed cross-architecture severity spectrum was largely an artifact of the flawed selection procedure rather than a property of end-to-end conformal risk training on ADFTD.

C.4 Subject-Level Tail-Risk Control

Subject-level CVaR reveals a complementary tail-risk–efficiency trade-off. On PTB at $\alpha = .20$, decreasing β from 1 to .7, .5, and .3 changes FNR/FPR from $.178/.264$ to $.113/.421$, $.078/.538$, and $.006/.947$, while the corresponding empirical subject-level CVaR decreases from .178 to .159, .155, and .017. On ADFTD, coverage/set size changes from $.848/2.010$ to $.913/2.383$, $.956/2.709$, and $1.000/3.000$, while the empirical CVaR of subject-level miscoverage decreases from .152 to .117, .083, and .000. Across both datasets, the transformed-risk estimates $\widehat{g}_t$ remain no smaller than the corresponding empirical CVaR values, consistent with the transformed-risk upper-bound construction. Thus, stronger tail protection reduces realized upper-tail subject risk at the cost of increasingly conservative predictions.

C.4.1 APAVA

APAVA is already in the trivial regime under subject-level mean-risk control. The same behavior persists under CVaR. For every tested combination

$$\alpha \in \{.05, .10, .20, .30\}, \qquad \beta \in \{.3, .5, .7\},$$

the calibrated predictor reduces to the always-positive rule, with FNR/FPR $= 0/1$.

Thus, strengthening the risk functional from the mean to CVaR cannot recover a nontrivial predictor when the effective number of calibration subjects is already too small. APAVA is therefore dominated by the finite-sample constraint rather than by the particular choice of tail fraction.

C.4.2 PTB

Table 14 empirically quantifies the subject-level tail-risk–efficiency trade-off on PTB. At $\alpha = .20$, decreasing β from 1 to .7, .5, and .3 changes FNR/FPR from $.178/.264$ to $.113/.421$, $.078/.538$, and $.006/.947$, respectively. The corresponding empirical subject-level CVaR decreases from .178 to .159, .155, and .017. Across these settings, the empirical transformed risk $\widehat{g}_t$ remains no smaller than the corresponding empirical CVaR, consistent with the transformed-risk upper-bound construction.

At $\alpha = .20$, stronger tail protection reduces realized upper-tail FNR at the cost of a higher false-positive rate. More broadly, PTB exhibits the expected tail-risk–efficiency trade-off, although the empirical CVaR is not strictly monotonic in β for every target level.

C.4.3 ADFTD

Table 15 empirically illustrates the corresponding tail-risk–efficiency trade-off in the multiclass setting. At $\alpha = .10$, the settings $\beta = .3$ and $\beta = .5$ yield the trivial all-three-class prediction set, with

Table 14: Subject-level CVaR control on PTB. FNR/FPR are subject-averaged evaluation metrics; $\widehat{\mathrm{CVaR}}_\beta$ is the empirical CVaR of subject-level FNR, and $\widehat{g}_t$ is the empirical transformed risk evaluated at the training-selected t.

α	β	FNR/FPR	$\widehat{\mathrm{CVaR}}_\beta$	$\widehat{g}_t$
.05	.3	.000/1.000	.000	.000
.05	.5	.000/1.000	.000	.002
.05	.7	.000/1.000	.000	.000
.10	.3	.000/1.000	.000	.008
.10	.5	.015/.844	.030	.037
.10	.7	.059/.612	.084	.087
.20	.3	.006/.947	.017	.080
.20	.5	.078/.538	.155	.179
.20	.7	.113/.421	.159	.169
.30	.3	.067/.587	.210	.283
.30	.5	.102/.457	.201	.270
.30	.7	.164/.296	.232	.272

Table 15: Subject-level CVaR control on ADFTD. Coverage/set size are subject-averaged evaluation metrics; $\widehat{\mathrm{CVaR}}_\beta$ and $\widehat{g}_t$ are computed from subject-level miscoverage losses.

α	β	Coverage/Size	$\widehat{\mathrm{CVaR}}_\beta$	$\widehat{g}_t$
.10	.3	1.000/3.000	.000	.017
.10	.5	1.000/3.000	.000	.021
.10	.7	.981/2.906	.026	.045
.20	.3	1.000/3.000	.000	.067
.20	.5	.956/2.709	.083	.128
.20	.7	.913/2.383	.117	.141
.30	.3	.960/2.738	.119	.193
.30	.5	.900/2.296	.183	.230
.30	.7	.847/2.009	.205	.245

coverage 1.000, average set size 3.000, and empirical subject-level miscoverage CVaR equal to 0. Increasing β to .7 yields a nontrivial operating point with coverage $.981$, average set size 2.906, and empirical CVaR $.026$.

Thus, relaxing the tail constraint permits smaller prediction sets but increases realized upper-tail miscoverage. Across the reported settings, the empirical transformed risk $\widehat{g}_t$ remains no smaller than the corresponding empirical CVaR, consistent with the transformed-risk upper-bound construction. Similar tail-risk–efficiency behavior is observed at $\alpha = .20$ and $.30$.

Compared with PTB, ADFTD exhibits a sharper saturation effect: sufficiently stringent settings require predicting all three classes. Relaxing the target risk α or increasing the tail fraction β allows nontrivial prediction sets at the cost of higher realized subject-level tail risk.

D Additional Methodological Details

This appendix provides derivations and implementation details supporting the constructions introduced in Section 4. We first make explicit the closed-form calibration rule for the classification losses used in our experiments, then describe differentiation through the calibrated threshold, the smooth efficiency objectives used for E2E training, and the corresponding subject-level mean- and tail-risk constructions.

D.1 Classification-Specific CRC and Closed-Form Calibration

For binary disease detection, let $p_\theta(x) \in [0,1]$ denote the predicted positive-class probability. Given a threshold λ,

$$\hat{y}_\lambda(x) = \mathbf{1}[p_\theta(x) \geq \lambda]\,. \tag{31}$$

For a disease-positive example, the false-negative loss is

$$\ell^{\mathrm{FN}}(x;\lambda) = \mathbf{1}[p_\theta(x) < \lambda]\,. \tag{32}$$

This loss is bounded in $[0,1]$ and nondecreasing in λ, so increasing the threshold trades a higher false-negative risk for fewer false positives.

For multiclass classification, we use the prediction set

$$C_\lambda(x) = \{k : p_\theta(x)_k \geq \lambda\}\,, \tag{33}$$

with miscoverage loss

$$\ell^{\mathrm{cov}}(x,y;\lambda) = \mathbf{1}[p_\theta(x)_y < \lambda]\,. \tag{34}$$

Again, the loss is bounded and nondecreasing in λ.

Both cases can therefore be written using a scalar calibration value

$$u_i = \begin{cases} p_\theta(x_i), & \text{binary FNR control,} \\ p_\theta(x_i)_{y_i}, & \text{multiclass coverage.} \end{cases} \tag{35}$$

Equivalently, define the nonconformity score

$$s_i = 1 - u_i. \tag{36}$$

Let

$$k_\alpha = \lceil (n+1)(1-\alpha) \rceil\,, \tag{37}$$

and let $s_{(1)} \leq \cdots \leq s_{(n)}$ denote the ordered calibration scores. Whenever $k_\alpha \leq n$, CRC reduces to the order-statistic rule

$$\hat{\lambda} = 1 - s_{(k_\alpha)}. \tag{38}$$

Thus, calibration requires no numerical optimization: the deployment threshold is obtained directly from one order statistic of the calibration probabilities.

If $k_\alpha > n$, the requested risk level cannot be certified with the available calibration sample under the finite-sample correction. In this case, we use the least aggressive admissible rule,

$$\hat{\lambda} = \lambda_{\min} = 0. \tag{39}$$

D.2 Differentiating Through the Calibrated Threshold

During E2E training, the calibrated threshold depends on the model parameters through the calibration probabilities. Away from ties, the selected order statistic corresponds locally to one calibration example,

$$\hat{\lambda}(\theta) = u_{i^\star}(\theta). \tag{40}$$

Although the discrete index $i^\star$ changes when the ordering changes, it is locally constant almost everywhere. The selected probability itself remains differentiable, giving

$$\frac{d\hat{\lambda}}{d\theta} = \frac{du_{i^\star}(\theta)}{d\theta} \tag{41}$$

almost everywhere.

Consequently, standard sorting and indexing operations retain the gradient of the selected calibration probability without requiring a custom backward procedure. Nondifferentiable points arise only at ties or order changes and form the usual measure-zero boundaries of the piecewise-smooth computation.

D.3 Differentiable Efficiency Objectives

The calibrated threshold enforces the risk constraint, while E2E training optimizes predictive efficiency on a separate pseudo-prediction subset.

For binary classification, efficiency is measured by FPR. We use the smooth surrogate

$$\widetilde{\mathcal{L}}_{\mathrm{FPR}} = \frac{1}{|\mathcal{B}^-|} \sum_{i \in \mathcal{B}^-} \sigma\left(\frac{p_\theta(x_i) - \hat{\lambda}(\theta)}{T}\right), \tag{42}$$

where $\mathcal{B}^-$ contains negative-class examples.

For multiclass prediction sets, efficiency is measured by set size. Because the true class is required for coverage, we optimize the inclusion of non-true classes:

$$\widetilde{\mathcal{L}}_{\mathrm{size}} = \frac{1}{|\mathcal{B}|(K-1)} \sum_{i \in \mathcal{B}} \sum_{k \neq y_i} \sigma\left(\frac{p_\theta(x_i)_k - \hat{\lambda}(\theta)}{T}\right). \tag{43}$$

The temperature $T > 0$ determines the smoothness of the surrogate. Smaller values more closely approximate the hard deployment rule, whereas larger values yield smoother gradients. The temperature-sensitivity experiments in Appendix C.3 examine the practical consequences of this trade-off.

D.4 Subject-Level CRC

For multiclass coverage, all calibration subjects constitute the exchangeable units. For binary FNR control, only disease-positive calibration subjects contribute to the target risk. Let $\mathcal{I}_+$ denote the set of disease-positive calibration subjects and let

$$S_+ = |\mathcal{I}_+|. \tag{44}$$

For notational convenience, define

$$\mathcal{I}_{\mathrm{eff}} = \begin{cases} \{1, \ldots, S\}, & \text{multiclass coverage,} \\ \mathcal{I}_+, & \text{binary FNR control,} \end{cases} \tag{45}$$

and

$$S_{\mathrm{eff}} = |\mathcal{I}_{\mathrm{eff}}| = \begin{cases} S, & \text{multiclass coverage,} \\ S_+, & \text{binary FNR control.} \end{cases} \tag{46}$$

For each $i \in \mathcal{I}_{\mathrm{eff}}$, let $\mathcal{J}_{i,\mathrm{eff}}$ denote the windows relevant to the target risk:

$$\mathcal{J}_{i,\mathrm{eff}} = \begin{cases} \{1, \ldots, K_i\}, & \text{multiclass coverage,} \\ \{j : y_{ij} = 1\}, & \text{binary FNR control,} \end{cases} \tag{47}$$

and let $K_{i,\mathrm{eff}} = |\mathcal{J}_{i,\mathrm{eff}}|$. The subject-level loss is then

$$L_i(\lambda) = \frac{1}{K_{i,\mathrm{eff}}} \sum_{j \in \mathcal{J}_{i,\mathrm{eff}}} \ell_{ij}(\lambda). \tag{48}$$

Thus, arbitrary within-subject dependence is absorbed into a single subject-level calibration unit.

The distinction between window- and subject-level calibration is then explicit. Pooling the relevant windows estimates

$$\widehat{R}_{\mathrm{win}}(\lambda) = \frac{\sum_{i \in \mathcal{I}_{\mathrm{eff}}} K_{i,\mathrm{eff}} L_i(\lambda)}{\sum_{i \in \mathcal{I}_{\mathrm{eff}}} K_{i,\mathrm{eff}}}, \tag{49}$$

so subjects contribute in proportion to the number of relevant windows they provide.

In contrast, subject-aware calibration estimates

$$\widehat{R}_{\mathrm{subj}}(\lambda) = \frac{1}{S_{\mathrm{eff}}} \sum_{i \in \mathcal{I}_{\mathrm{eff}}} L_i(\lambda), \tag{50}$$

giving equal weight to each exchangeable subject relevant to the target risk.

Applying CRC directly across the S_{eff} subject-level losses yields

$$\hat{\lambda}_{\text{subj}} = \max\left\{\lambda_{\min}, \sup\left\{\lambda : \frac{1 + \sum_{i \in \mathcal{I}_{\text{eff}}} L_i(\lambda)}{S_{\text{eff}} + 1} \leq \alpha\right\}\right\}. \tag{51}$$

Under subject-level exchangeability,

$$\mathbb{E}\left[L_{\text{new}}(\hat{\lambda}_{\text{subj}})\right] \leq \alpha. \tag{52}$$

The finite-sample correction therefore depends on the number of exchangeable subjects relevant to the target risk, S_{eff}, rather than on the nominal number of windows. In particular, $S_{\text{eff}} = S$ for multiclass coverage, whereas $S_{\text{eff}} = S_+$ for binary FNR control. Additional windows can improve estimation of each subject-specific loss $L_i(\lambda)$, but they do not create additional exchangeable calibration units. This is the statistical source of the nominal sample-size illusion discussed in the main text.

D.5 Subject-Level CVaR Control

Mean subject-level CRC controls the expected loss of a newly sampled subject, but does not directly target subjects in the upper tail of the loss distribution. To obtain tail-sensitive control, let $L_i(\lambda) \in [0, 1]$ denote the subject-level loss and let $\beta \in (0, 1]$ denote the protected tail fraction.

For a fixed auxiliary threshold t, define the transformed loss

$$\widetilde{L}_{i,t}(\lambda) = t + \frac{1}{\beta}\left(L_i(\lambda) - t\right)_+. \tag{53}$$

Since $L_i(\lambda) \in [0, 1]$, a pointwise upper bound is

$$B_t = t + \frac{1}{\beta}(1 - t)_+. \tag{54}$$

For fixed t, the transformed loss remains bounded and nondecreasing in λ. CRC can therefore be applied directly:

$$\hat{\lambda}_{\text{CVaR}} = \max\left\{\lambda_{\min}, \sup\left\{\lambda : \frac{B_t + \sum_{i=1}^{S} \widetilde{L}_{i,t}(\lambda)}{S + 1} \leq \alpha\right\}\right\}. \tag{55}$$

To connect this construction to CVaR, define

$$g_t(\lambda) = t + \frac{1}{\beta}\mathbb{E}\left[(L(\lambda) - t)_+\right]. \tag{56}$$

By the Rockafellar–Uryasev representation,

$$\text{CVaR}_\beta\big(L(\lambda)\big) = \min_{t'} g_{t'}(\lambda) \leq g_t(\lambda). \tag{57}$$

Therefore, controlling the expectation of $\widetilde{L}_{i,t}(\lambda)$ for any calibration-independent t also controls an upper bound on subject-level CVaR.

Why t is selected before calibration. The auxiliary threshold t must be chosen independently of the subjects used for conformal calibration. Selecting t on those same calibration subjects would introduce an additional data-dependent optimization step, so the direct finite-sample CRC argument above would no longer apply without further analysis.

We therefore select t using training data only and keep it fixed before conformal calibration. When $\beta = 1$, choosing $t = 0$ recovers the subject-level mean-risk construction in Eq. 51.

D.6 Fresh Calibration after End-to-End Training

The calibration performed inside E2E optimization is used only to construct the training objective. At each optimization step, a minibatch is partitioned into pseudo-calibration and pseudo-prediction subsets. The former determines the differentiable threshold $\hat{\lambda}(\theta)$, while the latter evaluates the corresponding efficiency surrogate.

These pseudo-calibration thresholds are not deployment thresholds and are discarded after optimization. Once training is complete, the final model is frozen and CRC is applied again using a fresh held-out calibration set.

This separation is important: pseudo-calibration determines how gradients are constructed during training, whereas the final held-out calibration step determines the deployed risk-controlling threshold.