Act or Defer: Error-Controlled Decision Policies for EHR Foundation Models

Abstract

Clinical deployment of foundation models requires decision policies that operate under explicit error budgets, such as a cap on false-positive clinical calls. Strong average accuracy alone does not guarantee safety: errors can concentrate among patients selected for action, leading to harm and inefficient use of healthcare resources. Here we develop STRATCP, a stratified conformal framework that turns foundation model predictions into decision-ready outputs through error-controlled selection and calibrated deferral. STRATCP first selects a subset of patients for immediate clinical action while controlling false discovery rates at a user-specified level. For the remaining patients, it returns prediction sets that achieve target coverage conditional on deferral, supporting confirmatory testing or expert review. We evaluate STRATCP on Electronic Health Record (EHR) foundation model predictions across EHRSHOT tasks spanning operational outcomes, assignment of new diagnoses, and anticipating lab test results, including length of stay, pancreatic cancer, and thrombocytopenia severity prediction. Across tasks, STRATCP controls class-specific false discovery rates among selected predictions and provides valid, selection-conditional coverage for deferred patients, with the largest gains on rarer and higher-stakes classes. STRATCP establishes error-controlled decision policies for safe deployment of medical foundation models.

1. Introduction

Electronic health record (EHR) foundation models (FMs) are increasingly used as general-purpose predictive engines for structured clinical data, spanning operational outcome prediction, diagnosis prediction, and longitudinal patient modeling (Renc et al., 2024; Fallahpour et al., 2024; Makarov et al., 2025; Kraljevic et al., 2024; Jiang et al., 2023; Waxler et al., 2025; Shmatko et al., 2025). Yet deployment requires more than strong average discrimination: clinicians need to know when a model prediction is reliable enough to support action, when it should be deferred, and what information can guide follow-up evaluation for deferred cases (Kompa et al., 2021; Ehrmann et al., 2023).

Conformal prediction (CP) provides distribution-free prediction sets with finite-sample coverage guarantees (Vovk et al., 2005; Angelopoulos & Bates, 2021), making it an attractive post-hoc calibration tool for foundation models. This has motivated the use of CP in clinical decision support (Vazquez & Facelli, 2022; Olsson et al., 2022; Sreenivasan et al., 2025; Shen et al., 2025) and in recent work on identifying high-quality foundation model outputs with formal guarantees (Gui et al., 2024). However, standard CP gives a marginal guarantee over the overall patient population. This does not directly control the error rate among the subset of patients selected for action, where clinical risk and resource use are concentrated. In deployment, a model may achieve nominal marginal coverage while still making unreliable high-confidence calls on harder, rarer, or more clinically consequential classes.

We introduce STRATCP, a stratified conformal framework that converts fixed EHR FM outputs into an act-or-defer decision policy (Fig. 1). In the action arm, STRATCP selects patients for immediate use while controlling class-specific false discovery rates among selected predictions. In the deferral arm, it returns prediction sets with selection-conditional coverage for patients not selected by the action arm. We evaluate STRATCP on EHRSHOT tasks spanning length of stay, pancreatic cancer, and thrombocytopenia severity, and show that it better tracks target selected coverage than standard CP and thresholding baselines while preserving calibrated uncertainty for deferred patients, with the largest gains on rarer and higher-stakes classes, such as pancreatic cancer and severe thrombocytopenia.

2. Method

2.1. Problem setup

Let $X \in \mathcal{X}$ denote a patient's structured EHR representation and $Y \in \mathcal{Y}$ the target label, with $\mathcal{Y} = \{0, 1\}$ for binary tasks and $\mathcal{Y} = \{1, \ldots, K\}$ for multiclass tasks. Following EHRSHOT, we use CLMBR-T-base, a 141M-parameter autoregressive EHR foundation model, with a task-specific

.**AUTHORERR: Missing \icmlcorrespondingauthor.**

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

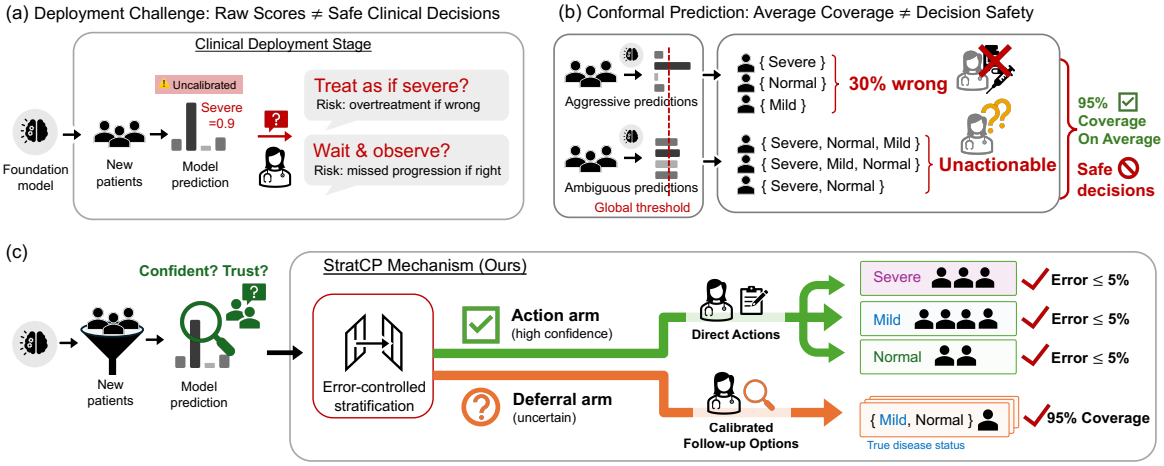

Figure 1. **From raw foundation model scores to STRATCP: error-controlled action and deferral.** (a) Raw model scores do not by themselves indicate whether a prediction is reliable enough for downstream use. (b) Marginal coverage from standard conformal prediction does not guarantee safety among patients selected for action. (c) STRATCP converts foundation model outputs into an action arm, which controls class-specific false discovery rates among selected predictions, and a deferral arm, which returns calibrated prediction sets with selection-conditional coverage for deferred patients.

logistic regression head and a frozen backbone throughout (Wornow et al., 2023). The resulting model outputs class probabilities

$$f(X) = \big(f(X,1),\dots,f(X,K)\big) \in \Delta^K. \tag{1}$$

Given a labeled calibration set and a held-out evaluation set, our goal is not merely to return the top-1 prediction, but to decide whether that prediction is reliable enough for direct downstream use under a pre-specified error budget. The calibration set is used only for post-hoc selection and uncertainty calibration; no retraining of the base model is performed.

2.2. Action arm: selecting predictions under an error budget

For each predicted class k, STRATCP considers only patients whose top-1 prediction is k and uses the class probability $f(x,k)$ as the confidence score. For a test patient j with top-1 prediction k, let $\tau_j = f(X_{n+j},k)$ be its confidence score, and let G_k^{cal} denote calibration patients whose top-1 prediction is also k. Let $E_i = \mathbf{1}\{\arg\max_{y'} f(X_i,y') \neq Y_i\}$ denote whether calibration example i is incorrectly classified. Following conformal selection (Jin & Candès, 2023), STRATCP computes the class-specific p-value

$$p_j = \frac{1 + \sum_{i\in G_k^{\text{cal}}} E_i \mathbf{1}\{f(X_i,k) \geq \tau_j\}}{1 + |G_k^{\text{cal}}|}. \tag{2}$$

This p-value is small when few calibration examples with the same predicted class are both incorrect and at least as confident as the test prediction. The resulting class-specific p-values are passed to the Benjamini–Hochberg procedure, yielding a selected subset of predictions with class-specific false discovery rate controlled at target level α. Equivalently, selected predictions have coverage at least $1-\alpha$ in expectation. Full algorithmic details and finite-sample guarantees are provided in Appendix Section A.2 and Section A.4.

This class-specific selection is important in EHR tasks with class imbalance or asymmetric difficulty: a model may be reliable on common classes while remaining unreliable on rarer or higher-stakes classes. STRATCP makes this heterogeneity explicit rather than averaging it away.

2.3. Deferral arm

Patients not selected for action are deferred. For these cases, STRATCP returns conformal prediction sets with selection-conditional validity (Jin & Ren, 2025). Let $V(x,y)$ denote a conformity score. Rather than calibrating against all calibration patients, STRATCP uses only calibration patients that would also have been deferred under the same action-arm rule. The deferred prediction set takes the form

$$\widehat{C}_{n+j} = \{y \in \mathcal{Y} : V(X_{n+j},y) \leq \widehat{q}_{1-\alpha,j}(y)\}, \tag{3}$$

where $\widehat{q}_{1-\alpha,j}(y)$ is the corresponding deferred-reference quantile. This yields selection-conditional coverage among deferred patients: after abstention, the returned prediction set contains the true label at the target rate. Full details are given in Appendix Section A.2 and Section A.4.

2.4. Baselines and metrics

We compare STRATCP against three post-hoc decision rules applied to the same fixed model outputs. First, standard CP constructs adaptive conformal prediction sets (Romano et al.,

2020); singleton sets are treated as actionable predictions, while non-singleton sets are treated as deferrals. This baseline provides marginal coverage over the overall population but does not directly target validity among the selected singleton predictions. Second, `calibrated threshold` uses the maximum predicted probability $\max_k f(X, k)$ as a confidence score and selects top-1 predictions whose score is at least a calibration-set threshold chosen to satisfy the target empirical error level. This represents a common high-confidence-only abstention baseline. Third, `cumulative threshold` sorts classes by predicted probability and returns the smallest prefix whose cumulative probability exceeds $1 - \alpha$; singleton prefixes are treated as actionable and larger sets as deferrals. This provides a simple probability-mass baseline that relies directly on the model's predicted class probabilities. Additional implementation details are provided in Section B.

Our primary metric is *class-specific selected coverage*, the empirical precision among selected predictions of each class. This evaluates the action arm: among patients selected for an actionable class-k prediction, it measures how often the prediction is correct. We report this metric separately by predicted class because errors can concentrate on rarer or harder classes despite acceptable aggregate performance.

We also report *deferred-set coverage*, which evaluates whether prediction sets returned for abstained patients contain the true label, and *selection efficiency*, measured as the number of patients selected for action at each target error level. Validity is the primary requirement; among methods that meet the target, higher selection efficiency indicates better use of the available error budget. Additional implementation details for all baselines and metrics are provided in Section B.

3. Experimental setup

We evaluate a frozen EHR foundation model on tasks drawn from the EHRSHOT benchmark (Wornow et al., 2023). We report three representative task families chosen to span operational, diagnostic, and laboratory prediction settings across both binary and multiclass outcomes, and for which the frozen EHR FM exhibits sufficient predictive signal to make the action–deferral tradeoff informative: length of stay (LOS), defined as whether total hospital stay is at least 7 days; pancreatic cancer, defined as first diagnosis within 365 days post-discharge; and thrombocytopenia severity immediately before lab measurement, with classes normal (≥ 150), mild (100–150), moderate (50–100), and severe (< 50) $\times 10^9$/L. Additional data splitting, calibration, resampling, baseline, and metric details are provided in Section B.

4. Results

The frozen EHR foundation model has nontrivial predictive signal across all three tasks, with AUROC/AUPRC summarized in Appendix Table 1; however, our primary question is whether predictions selected for downstream use satisfy a deployment-time error budget.

Figure 2 shows selected coverage as a function of the target error level α across the three chosen task families. For length of stay, STRATCP remains close to the target selected coverage for both Short LOS and Long LOS: at $\alpha = 0.10$, selected coverage is 0.914 for Short LOS and 0.956 for Long LOS. The difference relative to `CP`, `Calibrated Threshold`, and `Cumulative Threshold` is more apparent on the Long LOS class, where the comparison methods fall below the target over a larger portion of the α range (Figure 2a). At $\alpha = 0.10$, STRATCP selects on average 457.3 Short LOS cases and 5.3 Long LOS cases, whereas CP selects 186.7 Short LOS cases and 115.6 Long LOS cases (Appendix Fig. 4). Thus, STRATCP remains efficient on the common class while abstaining much more on the uncertain Long LOS class in order to meet the selected-coverage target.

For pancreatic cancer, the separation between methods is concentrated on the positive class. At $\alpha = 0.10$, STRATCP attains selected coverage of 0.990 on the No pancreatic cancer class and 0.991 on the Pancreatic cancer class, whereas `CP` attains 0.990 and 0.107, respectively (Figure 2b). The corresponding average numbers selected are 577.1 and 0.2 for STRATCP, versus 270.1 and 79.4 for CP (Appendix Fig. 4). These values show that on the rare positive class, STRATCP largely abstains rather than issuing low-reliability selected predictions, while CP continues to select many positive cases at very low selected coverage.

For multiclass thrombocytopenia, STRATCP also remains closer to the target across the Normal, Mild, Moderate, and Severe classes. At $\alpha = 0.15$, selected coverage on the Severe class is 0.957 for STRATCP versus 0.669 for `CP` (Figure 2c). At $\alpha = 0.15$, STRATCP selects on average 1102.3 Normal cases and 1.3 Severe cases, whereas `CP` selects 980.7 Normal cases and 8.6 Severe cases (Appendix Fig. 4). These results show that STRATCP remains efficient on common classes while selecting very few cases on rarer and more consequential severity states unless they can be supported at the target error level.

4.1. Deferred-set coverage remains valid for deferred patients

Figure 3 shows coverage among deferred thrombocytopenia patients. Deferred-set coverage under STRATCP tracks the target across α, showing that STRATCP not only controls error among selected patients, but also returns well-calibrated

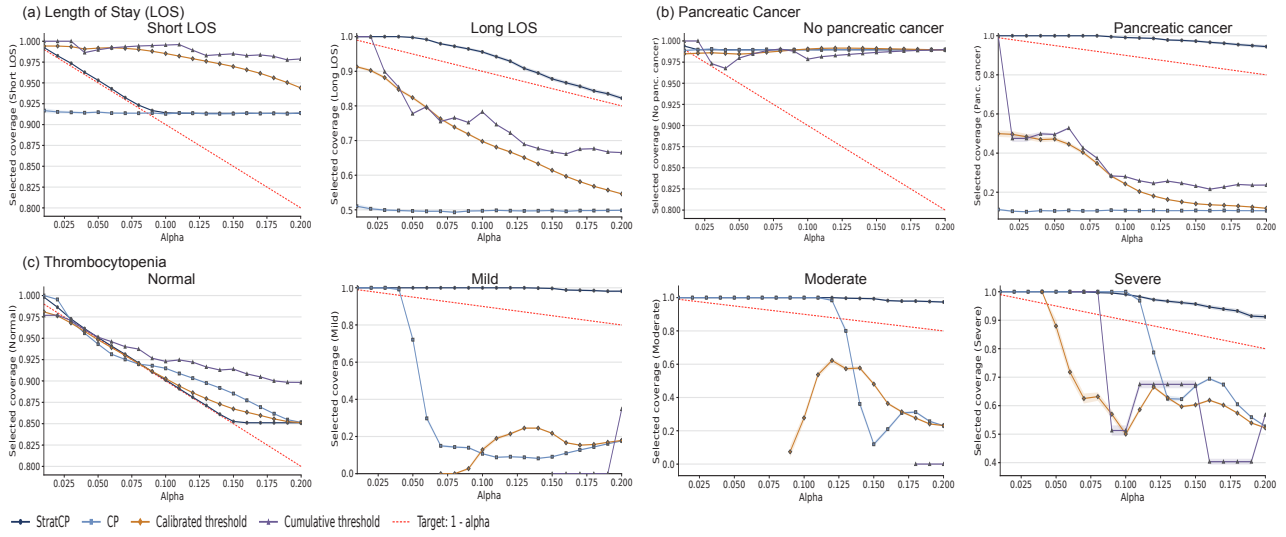

Figure 2. **Selected coverage across representative EHRSHOT tasks.** Selected coverage is plotted against the target error level α for three representative task families: (a) length of stay (Short LOS, Long LOS), (b) pancreatic cancer (No pancreatic cancer, Pancreatic cancer), and (c) thrombocytopenia (Normal, Mild, Moderate, Severe). The red dashed line denotes the target coverage $1 - \alpha$. Across tasks, STRATCP more consistently tracks the target selected coverage than `CP`, `Calibrated threshold`, and `Cumulative threshold`, with the largest gaps on harder positive or severity-specific classes.

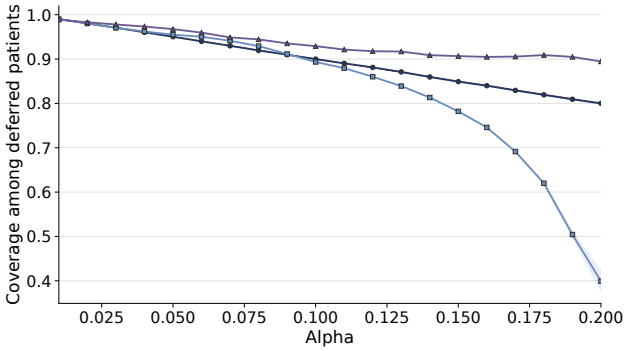

Figure 3. **Deferred-set coverage for thrombocytopenia.** Coverage among deferred patients is plotted against the target error level α. The dashed reference line denotes the target coverage $1 - \alpha$. STRATCP maintains valid deferred-set coverage, supporting calibrated follow-up prediction sets for patients not selected in the action arm.

prediction sets for patients it defers. Thus, deferred patients receive prediction sets with valid coverage rather than being simply rejected, complementing the action-arm results and supporting an act-or-defer decision workflow (deferred-coverage plots for the two binary tasks are provided in Appendix Fig. 5).

5. Discussion

This work shows that deployment of EHR foundation models requires action-aligned calibration, not only strong average predictive performance. Across length of stay, pancreatic cancer, and thrombocytopenia, STRATCP separates patients into an action arm with explicit error control and a deferral arm with calibrated follow-up uncertainty. The largest gains appear on harder, rarer, and higher-stakes classes, where low-reliability selected predictions are least acceptable.

A practical advantage of STRATCP is that it is model-agnostic and post hoc: it wraps fixed EHR foundation model outputs without retraining the backbone. Its guarantees still depend on exchangeability and adequate calibration data, especially for rare classes, and its selection efficiency is naturally limited by the predictive strength of the underlying model. As EHR foundation models improve, the same act-or-defer layer can translate stronger predictions into more actionable cases while preserving explicit error control. These results support error-controlled selection and calibrated deferral as a practical deployment layer for medical foundation models.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Ehrmann, D. E., Joshi, S., Goodfellow, S. D., Mazwi, M. L., and Eytan, D. Making machine learning matter to clinicians: model actionability in medical decision-making. *NPJ Digital Medicine*, 6(1):7, 2023.

Fallahpour, A., Alinoori, M., Ye, W., Cao, X., Afkanpour, A., and Krishnan, A. Ehrmamba: Towards generaliz-

able and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

Gui, Y., Jin, Y., and Ren, Z. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37: 73884–73919, 2024.

Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H. A., Laufer, I., Punjabi, P., et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969): 357–362, 2023.

Jin, Y. and Candès, E. J. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Jin, Y. and Ren, Z. Confidence on the focal: Conformal prediction with selection-conditional coverage. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf016, 2025.

Kompa, B., Snoek, J., and Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H., Yeung, J. A., Deng, A., Balston, A., Ross, J., Idowu, E., et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

Makarov, N., Bordukova, M., Quengdaeng, P., Garger, D., Rodriguez-Esteban, R., Schmich, F., and Menden, M. P. Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine*, 8(1):588, 2025.

Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., Delahunt, B., Lindskog, C., Janssen, E. A., Blilie, A., et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature communications*, 13 (1):7761, 2022.

Renc, P., Jia, Y., Samir, A. E., Was, J., Li, Q., Bates, D. W., and Sitek, A. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7(1):256, 2024.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Shen, Z., Chakraborti, T., Banerji, C. R., and Ding, X. Conformal prediction quantifies wearable cuffless blood pressure with certainty. *Scientific Reports*, 15(1):26697, 2025.

Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T., and Gerstung, M. Learning the natural history of human disease with generative transformers. *Nature*, pp. 1–9, 2025.

Sreenivasan, A. P., Vaivade, A., Noui, Y., Khoonsari, P. E., Burman, J., Spjuth, O., and Kultima, K. Conformal prediction enables disease course prediction and allows individualized diagnostic uncertainty in multiple sclerosis. *npj Digital Medicine*, 8(1):224, 2025.

Vazquez, J. and Facelli, J. C. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.

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

Waxler, S., Blazek, P., White, D., Sneider, D., Chung, K., Nagarathnam, M., Williams, P., Voeller, H., Wong, K., Swanhorst, M., et al. Generative medical event models improve with scale. *arXiv:2508.12104*, 2025.

Wornow, M., Thapa, R., Steinberg, E., Fries, J., and Shah, N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

A. Algorithmic Details and Finite-Sample Guarantees

This appendix provides the algorithmic details and finite-sample guarantees for the action and deferral arms of STRATCP. We specialize the presentation to the classification setting considered in the main paper.

A.1. Notation and assumptions

Let $X \in \mathcal{X}$ denote a patient's structured EHR representation and $Y \in \mathcal{Y} = \{1, \dots, K\}$ the corresponding label. For each input x, the fixed EHR foundation model and task-specific head output class probabilities $f(x) \in \Delta^K$, where $f(x, k)$ denotes the predicted probability of class k. Let

$$\widehat{y}(x) = \arg\max_{k \in \mathcal{Y}} f(x, k)$$

denote the top-1 predicted class.

We use a labeled calibration set

$$\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$$

and a held-out evaluation set

$$\mathcal{D}_{\text{test}} = \{(X_{n+j}, Y_{n+j})\}_{j=1}^{m}.$$

At inference time, the procedure uses $\mathcal{D}_{\text{cal}}$ and the test features $\{X_{n+j}\}_{j=1}^{m}$; the test labels are used only for evaluation.

The guarantees below are stated conditional on the fitted model f. They require that the calibration and evaluation examples are exchangeable after conditioning on f and any fixed preprocessing choices. In our experiments, the EHR foundation model and task-specific heads are trained before the calibration/evaluation split considered here; the calibration and evaluation subsets are not used for model training, hyperparameter selection, or in-training validation. During this study, these subsets are used only for post-hoc calibration, selection, prediction-set construction, and final evaluation.

A.2. Action-arm conformal selection

The action arm selects a subset of patients for which the top-1 model prediction is reliable enough for direct downstream use. Because reliability may differ substantially across predicted classes, STRATCP performs selection separately within each predicted class.

For a predicted class k, define the class-specific correctness criterion

$$\mathcal{I}_k(f, x, y) = \mathbf{1}\{\widehat{y}(x) = y = k\},$$

which indicates that the top-1 prediction is class k and is correct. The corresponding eligibility group consists of patients whose top-1 prediction is class k:

$$G_k^{\text{cal}} = \{i \in [n] : \widehat{y}(X_i) = k\}, \qquad G_k^{\text{test}} = \{j \in [m] : \widehat{y}(X_{n+j}) = k\}.$$

In our experiments, the score for class-k selection is the predicted class probability

$$s_k(x) = f(x, k).$$

Larger values of $s_k(x)$ indicate stronger model confidence in the class-k prediction.

For each calibration example $i \in G_k^{\text{cal}}$, let

$$E_i = \mathbf{1}\{\widehat{y}(X_i) \neq Y_i\}$$

denote whether the top-1 prediction is incorrect. For each test example $j \in G_k^{\text{test}}$, STRATCP computes the conformal p-value

$$p_j = \frac{1 + \sum_{i \in G_k^{\text{cal}}} E_i \, \mathbf{1}\{s_k(X_i) \geq s_k(X_{n+j})\}}{1 + |G_k^{\text{cal}}|}.$$

This p-value is small when few same-predicted-class calibration examples are both incorrect and at least as confident as the test example.

The class-specific selected set is obtained by applying the Benjamini–Hochberg procedure to the p-values $\{p_j : j \in G_k^{\text{test}}\}$. Specifically, let $p_{(1)} \le \cdots \le p_{(|G_k^{\text{test}}|)}$ denote the sorted p-values and set

$$\widehat{\ell}_k = \max\left\{\ell : p_{(\ell)} \le \frac{\alpha \ell}{|G_k^{\text{test}}|}\right\},$$

with $\widehat{\ell}_k = 0$ if the set is empty. The selected class-k patients are

$$\mathcal{R}_k = \left\{j \in G_k^{\text{test}} : p_j \le \frac{\alpha \widehat{\ell}_k}{|G_k^{\text{test}}|}\right\}.$$

The full selected set is $\mathcal{R} = \cup_{k=1}^{K} \mathcal{R}_k$. Patients in $\mathcal{R}$ are assigned their top-1 predicted label; patients not in $\mathcal{R}$ are deferred to the deferral arm.

The procedure can be summarized as follows.

1. For each class k, form the calibration and test eligibility groups G_k^{cal} and G_k^{test} based on the top-1 predicted class.
2. For each $i \in G_k^{\text{cal}}$, compute the calibration error indicator $E_i = \mathbf{1}\{\widehat{y}(X_i) \neq Y_i\}$.
3. For each $j \in G_k^{\text{test}}$, compute the conformal p-value p_j using same-predicted-class calibration examples.
4. Apply the Benjamini–Hochberg procedure to $\{p_j : j \in G_k^{\text{test}}\}$ at target level α.
5. Select the rejected test examples as $\mathcal{R}_k$ and assign them the singleton prediction $\widehat{y}(X_{n+j}) = k$.

This class-specific construction directly matches the selected-coverage curves reported in the main text: for each class k, the false discovery proportion is one minus the empirical precision among selected predictions of class k.

A.3. Deferral-arm conformal prediction

Patients not selected by the action arm are deferred. For each deferred patient, STRATCP returns a conformal prediction set with coverage conditional on deferral. This ensures that abstention is accompanied by calibrated follow-up uncertainty rather than by an uninformative rejection.

Let $V(x, y)$ denote a conformity score, where smaller values indicate that label y is more compatible with input x. In our classification experiments, we use the adaptive prediction set (APS) score

$$V(x, y) = \sum_{y' \in \mathcal{Y}:\, f(x,y') \ge f(x,y)} f(x, y').$$

This score orders labels by decreasing predicted probability and assigns each candidate label the cumulative probability mass up to that label.

The key step in the deferral arm is to calibrate only against examples that are compatible with the same deferral event. Let $\mathcal{A}$ denote the full action-arm selection rule described above. For a deferred test patient $j \notin \mathcal{R}$ and a candidate label $y \in \mathcal{Y}$, define the selection-aware reference set

$$\mathcal{D}_{n+j}(y) = \left\{i \in [n] : j \notin \mathcal{A}\Big(\mathcal{D}_{\text{cal}}^{\text{swap}(i,j)}(y), \{X_{n+\ell}\}_{\ell=1}^{m}\Big)\right\}.$$

Here, $\mathcal{D}_{\text{cal}}^{\text{swap}(i,j)}(y)$ denotes the calibration set obtained by replacing the i-th calibration example with the hypothesized test example (X_{n+j}, y), while the original calibration example (X_i, Y_i) is treated as occupying the corresponding test position in the selection calculation. Intuitively, $\mathcal{D}_{n+j}(y)$ contains calibration examples that remain comparable to test patient j after conditioning on the same deferral decision.

For each candidate label y, define the deferred-reference quantile

$$\widehat{q}_{1-\alpha,j}(y) = \text{Quantile}_{1-\alpha}\left(\{V(X_i, Y_i) : i \in \mathcal{D}_{n+j}(y)\} \cup \{+\infty\}\right),$$

where the additional $+\infty$ term is the standard conformal finite-sample correction. The deferred prediction set is then

$$\widehat{C}_{n+j} = \{y \in \mathcal{Y} : V(X_{n+j}, y) \leq \widehat{q}_{1-\alpha,j}(y)\}.$$

The deferral-arm construction can be summarized as follows.

1. Run the action arm and identify the deferred patients $\{j : j \notin \mathcal{R}\}$.
2. For each deferred test patient j and each candidate label $y \in \mathcal{Y}$, form the selection-aware reference set $\mathcal{D}_{n+j}(y)$ by identifying calibration examples that would remain comparable to j under the same deferral event.
3. Compute the deferred-reference quantile $\widehat{q}_{1-\alpha,j}(y)$ from the conformity scores of examples in $\mathcal{D}_{n+j}(y)$.
4. Include label y in $\widehat{C}_{n+j}$ if $V(X_{n+j}, y) \leq \widehat{q}_{1-\alpha,j}(y)$.

For the classification tasks considered in this paper, the label space is small enough that this construction can be implemented by evaluating the action-arm rule under the required label swaps.

A.4. Finite-sample guarantees

We now state the finite-sample guarantees corresponding to the action and deferral arms. Both guarantees hold conditional on the fitted model f and rely only on exchangeability between calibration and evaluation examples.

Proposition A.1 (Class-specific FDR control)**.** *Assume the calibration and evaluation examples are exchangeable and that f is fixed independently of them. For any predicted class k and any fixed score function s_k, the action-arm selected set $\mathcal{R}_k$ satisfies*

$$\mathbb{E}\left[\frac{\sum_{j\in\mathcal{R}_k} \mathbf{1}\{Y_{n+j} \neq k\}}{|\mathcal{R}_k| \vee 1}\right] \leq \alpha.$$

Thus, among patients selected for a class-k action, the expected fraction of incorrect predictions is at most α. Equivalently, the expected selected coverage for class-k predictions is at least $1-\alpha$, matching the class-specific selected coverage reported in the main text.

Proof sketch. For each predicted class k, the conformal p-value compares a test example to exchangeable calibration examples in the same predicted-class stratum. Under the null event that the class-k top-1 prediction is incorrect, this conformal p-value is valid. Applying the Benjamini–Hochberg procedure to these valid p-values controls the expected false discovery proportion among selected class-k predictions. The result follows from the conformal selection framework; see (Jin & Candès, 2023) for the full proof.

Proposition A.2 (Selection-conditional coverage for deferred patients)**.** *Under the same exchangeability and fixed-model assumptions, for any deferred test patient j and any predicted-class stratum k,*

$$\mathbb{P}\left(Y_{n+j} \in \widehat{C}_{n+j} \;\middle|\; j \in G_k^{\text{test}} \setminus \mathcal{R}_k\right) \geq 1-\alpha.$$

Therefore, conditional on the patient not being selected for immediate action, the deferred prediction set contains the true label with probability at least $1-\alpha$.

Proof sketch. The deferral event depends on the calibration data and the test features through the action-arm selection rule. The swap-based construction of $\mathcal{D}_{n+j}(y)$ identifies calibration examples that are exchangeable with the test example after conditioning on the same deferral event. Applying a conformal quantile to this selection-aware reference set yields finite-sample coverage for the deferred test patient. The additional $+\infty$ term gives the usual finite-sample conformal correction. This is an application of the joint Mondrian conformal inference framework for selection-conditional coverage; see (Jin & Ren, 2025) for the full proof.

B. Experimental and Implementation Details

B.1. EHRSHOT tasks and model outputs

We evaluate STRATCP on three representative tasks from EHRSHOT (Wornow et al., 2023), chosen to span qualitatively different clinical prediction settings and for which the frozen EHR FM exhibits sufficient predictive signal to make the action–deferral tradeoff informative.

Length of stay. Length of stay is a binary operational outcome task indicating whether a patient's total hospital stay is at least 7 days. This task represents a common deployment setting in which reliable predictions could support bed management, discharge planning, and early care coordination.

Pancreatic cancer. Pancreatic cancer is a binary assignment-of-new-diagnosis task indicating whether a patient receives a first pancreatic cancer diagnosis within 365 days after discharge. This task represents a rare and high-stakes prediction setting, where unreliable selected positive predictions could lead to unnecessary follow-up evaluation and resource use.

Thrombocytopenia. Thrombocytopenia is a multiclass lab-anticipation task defined by platelet-count severity immediately before the lab is measured. We use the clinically ordered severity labels normal (≥ 150), mild (100–150), moderate (50–100), and severe (< 50) $\times 10^9$/L. This task evaluates whether the act-or-defer policy remains valid in a multiclass setting with clinically meaningful severity states.

Following EHRSHOT, we use CLMBR-T-base, a 141M-parameter autoregressive EHR foundation model pretrained on longitudinal structured EHR data, with a task-specific logistic regression head for each downstream task (Wornow et al., 2023). The pretrained backbone is held fixed throughout. All conformal and baseline methods are applied post hoc to the same fixed class-probability outputs $f(X)$.

B.2. Calibration, evaluation, and resampling

For each task, we use the held-out test set designated by EHRSHOT and further split it at the patient level into calibration and evaluation subsets. Specifically, within the EHRSHOT held-out set, 40% of patients are assigned to calibration and 60% are assigned to evaluation. The calibration subset is used only for post-hoc calibration, selection, and threshold estimation; it is not used to retrain the EHR foundation model or the task-specific head. The evaluation subset is used only to compute selected coverage, deferred-set coverage, selection efficiency, AUROC, AUPRC, and class counts.

To account for variability from the calibration/evaluation split, we repeat this patient-level random splitting procedure over 500 Monte Carlo splits and report averaged results. For the selected-coverage, deferred-coverage, and selection-efficiency curves, the plotted line denotes the mean across Monte Carlo splits, and the shaded band denotes the standard error across splits.

All methods are evaluated on the same fixed model outputs, calibration subsets, and evaluation subsets for each split. For a given target error level α, each method produces either an actionable singleton prediction or a deferral output. For STRATCP, actionable predictions come from the action arm, while deferred patients receive conformal prediction sets from the deferral arm.

We evaluate methods over 20 evenly spaced target error levels from $\alpha = 0.01$ to $\alpha = 0.20$. For the base predictive performance in Appendix Table 1, standard errors were estimated using 2,000 bootstrap resamples of the test set. For binary tasks, we report standard AUROC and positive-class AUPRC. For thrombocytopenia, we report weighted one-vs-rest AUROC and weighted AUPRC.

B.3. Baseline implementation details

We compare STRATCP against three post-hoc decision rules applied to the same fixed model probabilities.

Standard conformal prediction. The standard CP baseline constructs adaptive prediction sets using the APS conformity score (Romano et al., 2020),

$$V(x, y) = \sum_{y' \in \mathcal{Y}:\, f(x,y') \geq f(x,y)} f(x, y').$$

The calibration set is used to compute the marginal conformal quantile at target coverage level $1-\alpha$. At evaluation time, singleton prediction sets are treated as actionable predictions, while non-singleton prediction sets are treated as deferrals. This baseline provides marginal coverage over the full evaluation population, but does not directly control the error rate among the subset of singleton predictions selected for action.

Calibrated maximum-probability threshold. The calibrated threshold baseline uses the maximum predicted probability

$$c(x) = \max_{k \in \mathcal{Y}} f(x, k)$$

as a confidence score and uses the top-1 class

$$\widehat{y}(x) = \arg\max_{k \in \mathcal{Y}} f(x, k)$$

as the candidate actionable label. On the calibration set, examples are sorted by decreasing $c(X_i)$, and tied confidence scores are grouped together so that a threshold at value t includes all examples with $c(X_i) \geq t$. For each target error level α, we choose the threshold that maximizes the number of selected calibration examples subject to the empirical selected-case error being at most α. If no threshold satisfies this constraint, the threshold is set to $+\infty$ and no test examples are selected. At evaluation time, patients with $c(X_{n+j})$ greater than or equal to the chosen threshold are selected for action and assigned their top-1 predicted label; the remaining patients are deferred. This represents a common high-confidence-only abstention baseline.

Cumulative-probability threshold. The cumulative threshold baseline sorts classes by predicted probability for each patient and returns the smallest prefix of labels whose cumulative probability mass exceeds $1-\alpha$. If this prefix contains a single label, the prediction is treated as actionable; if the prefix contains multiple labels, the patient is treated as deferred. This baseline uses the model's predicted probability mass directly, but unlike STRATCP it does not calibrate selected predictions to control selected-case error.

Common evaluation protocol. All baselines use the same calibration and evaluation splits as STRATCP. For every method, we compute class-specific selected coverage, deferred-set coverage, and selection efficiency at each target error level α.

B.4. Metric implementation details

We evaluate three deployment-oriented metrics: selected coverage, deferred-set coverage, and selection efficiency.

Let $A_j \in \{0, 1\}$ indicate whether evaluation patient j is selected for action, and let $\widehat{y}_j$ denote the actionable predicted label when $A_j = 1$. For methods that return prediction sets, singleton sets are treated as actionable predictions and non-singleton sets are treated as deferrals. For STRATCP, actionable predictions are the patients selected by the action arm.

Selected coverage. For each predicted class k, selected coverage is the empirical precision among selected predictions of class k:

$$\widehat{\mathrm{SelCov}}_k = \frac{\sum_{j=1}^{m} \mathbf{1}\{A_j = 1,\ \widehat{y}_j = k,\ Y_{n+j} = k\}}{\sum_{j=1}^{m} \mathbf{1}\{A_j = 1,\ \widehat{y}_j = k\}}.$$

This is the main evaluation metric because it directly measures whether actionable class-k predictions satisfy the target error budget. For binary tasks, we report selected coverage separately for the two predicted classes. For thrombocytopenia, we report selected coverage separately for the normal, mild, moderate, and severe classes. If no examples are selected for a given class at a given α, the corresponding selected coverage is undefined and is excluded when aggregating or plotting coverage performance metrics.

Deferred-set coverage. Let $D_j = 1 - A_j$ indicate that patient j is deferred, and let $\widehat{C}_{n+j}$ denote the prediction set returned for a deferred patient. Deferred-set coverage is

$$\widehat{\mathrm{Cov}}_{\mathrm{def}} = \frac{\sum_{j=1}^{m} \mathbf{1}\{D_j = 1,\ Y_{n+j} \in \widehat{C}_{n+j}\}}{\sum_{j=1}^{m} \mathbf{1}\{D_j = 1\}}.$$

Table 1. **Test-set predictive performance of the frozen EHR foundation model on the EHRSHOT tasks.** For binary tasks, we report the standard test AUROC and positive-class AUPRC; for thrombocytopenia, we report weighted one-vs-rest AUROC and weighted AUPRC. Standard errors (SEs) were estimated using 2,000 bootstrap resamples of the test set.

Task	Test AUROC (SE)	Test AUPRC (SE)	Outcome	Test class counts
Length of stay (LOS)	$0.840\,(\pm 0.012)$	$0.620\,(\pm 0.029)$	Binary	$n_0 = 936$, $n_1 = 302$
Pancreatic cancer	$0.797\,(\pm 0.046)$	$0.251\,(\pm 0.068)$	Binary	$n_0 = 1206$, $n_1 = 40$
Thrombocytopenia	$0.765\,(\pm 0.013)$	$0.808\,(\pm 0.010)$	Multiclass	$n_0 = 1645$, $n_1 = 213$, $n_2 = 90$, $n_3 = 50$

This metric evaluates the deferral arm: among patients not selected for direct action, it measures whether the returned prediction set contains the true label. If no patients are deferred at a given α, deferred-set coverage is undefined and is excluded when aggregating or plotting coverage performance metrics.

Selection efficiency. Selection efficiency is measured by the number of patients selected for action:

$$\widehat{N}_{\text{sel}} = \sum_{j=1}^{m} \mathbf{1}\{A_j = 1\}.$$

We also report class-specific selection counts,

$$\widehat{N}_{\text{sel},k} = \sum_{j=1}^{m} \mathbf{1}\{A_j = 1,\ \widehat{y}_j = k\}.$$

Validity is the primary requirement: among methods that achieve the target selected coverage, larger selected sets indicate better use of the available error budget.

Base predictive performance. For context, we also report AUROC and AUPRC for the frozen EHR foundation model. These metrics characterize the underlying predictive signal of the fixed model, but they do not by themselves determine whether selected predictions satisfy a deployment-time error budget.

C. Additional Results

C.1. Base predictive performance

Appendix Table 1 summarizes the test-set predictive performance of the frozen EHR foundation model on the three EHRSHOT tasks. These results provide context for the downstream act-or-defer evaluation: the model has nontrivial predictive signal across tasks, but AUROC and AUPRC do not directly evaluate whether selected predictions satisfy a user-specified error budget.

C.2. Selection efficiency

Appendix Fig. 4 reports the number of selected patients as a function of the target error level α. These curves complement the selected coverage results in the main text by showing how each method uses the available error budget.

Across tasks, STRATCP selects many patients for common and more reliable classes, such as Short LOS, No pancreatic cancer, and Normal thrombocytopenia. In contrast, it selects substantially fewer patients for rarer or more uncertain classes, such as Long LOS, pancreatic cancer, and higher-severity thrombocytopenia states. This behavior is consistent with the goal of error-controlled deployment: STRATCP acts when the evidence supports the target selected-coverage level and abstains when the prediction cannot be made reliably enough.

C.3. Deferred-set coverage for binary tasks

Appendix Fig. 5 shows deferred-set coverage for the two binary tasks, Length of Stay and Pancreatic Cancer. Together with the main-text thrombocytopenia deferred-coverage result, these plots show that STRATCP provides valid follow-up uncertainty quantification for patients not selected in the action arm.

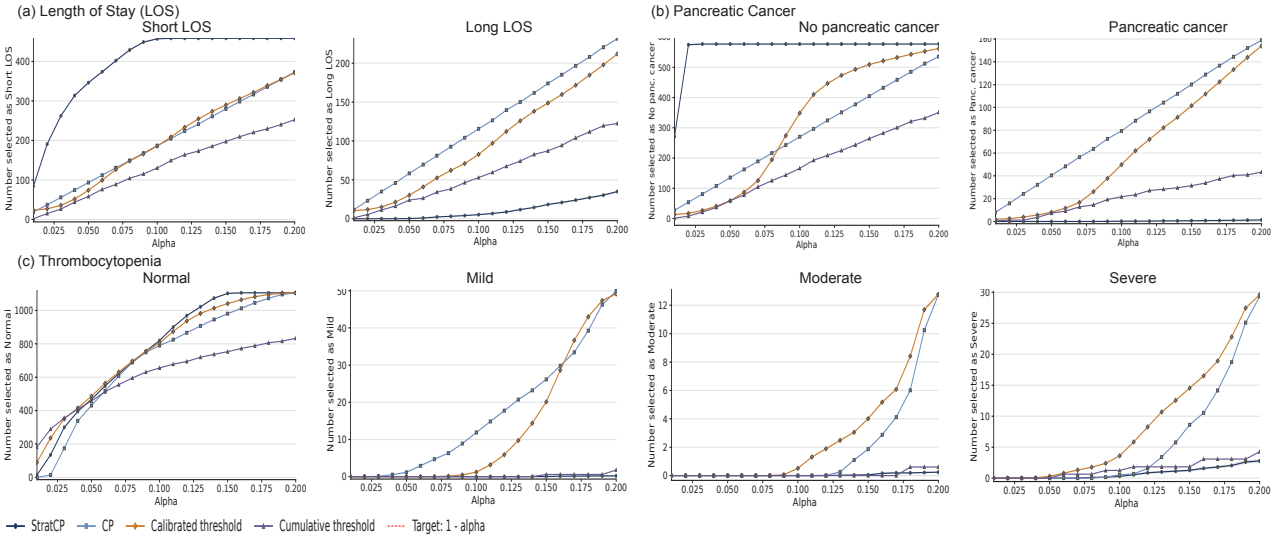

Figure 4. **Number of selected patients across three chosen EHR tasks.** For each task, class, and target error level α, curves show the number of patients selected by each method. Across tasks, STRATCP selects many patients on common classes (e.g., Short LOS, No pancreatic cancer, and Normal thrombocytopenia) while selecting substantially fewer patients on rarer or more uncertain classes (e.g., Long LOS, pancreatic cancer, and higher-severity thrombocytopenia states), consistent with uncertainty-driven abstention under an explicit error budget.

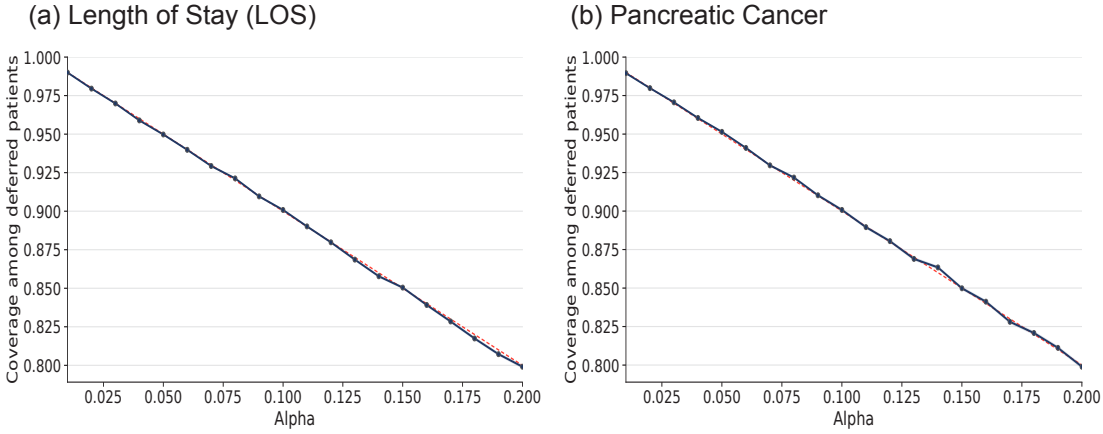

Figure 5. **Deferred-set coverage for the binary tasks.** Coverage among deferred patients is shown for (a) Length of Stay and (b) Pancreatic Cancer under STRATCP. In these binary settings, deferred-set coverage follows the target line $1 - \alpha$. Comparison methods are omitted because their binary selection rules yield full deferred coverage across all α.