SELF-CONSISTENT CONFORMAL PREDICTION: RELIABLE DECISIONS UNDER PERFORMATIVE FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction affords finite-sample marginal coverage under exchangeability, making it attractive for medical decision support. Yet in sequential care, when a prediction set contains clinically concerning outcomes, clinicians may be more likely to initiate preventive treatment or intensify an existing intervention, changing the very outcome the model aims to predict. This can break the exchangeability assumption underlying the coverage guarantee. A seemingly simple workaround is to ignore this feedback and calibrate under the historical treatment process. However, ignoring treatment feedback can lead to systematic coverage errors at deployment, so the nominal coverage level may no longer provide a reliable measure for clinical decision making. For all that, we show that this feedback can nevertheless be handled before deployment. We propose *Self-Consistent Conformal Prediction* (SCCP), a method for reliable uncertainty quantification in sequential care under performative feedback. The central difficulty is that the deployment distribution itself depends on the threshold being calibrated. We formulate this coupling as a self-consistent calibration problem, show that its stagewise coverage can be identified from historical trajectories, and use this result to construct a threshold schedule for the treatment process induced by its own use. We prove asymptotic stagewise marginal coverage without requiring post-deployment outcomes. Experiments on one synthetic benchmark and semi-synthetic treatment settings built from four clinical datasets show that SCCP maintains coverage close to the nominal level while retaining informative prediction sets.

1 INTRODUCTION

Interest has been growing in reliable methods for treatment decision-making (Azizi et al., 2026; Melnychuk et al., 2024). Clinicians must account for uncertainty, as different plausible clinical outcomes may warrant different interventions (Hess & Feuerriegel, 2025; Cortes-Gomez et al., 2025). Conformal prediction (CP) offers an elegant way to express this uncertainty through prediction sets. Under exchangeability, these sets afford finite-sample coverage of the true outcome with probability no less than $1 - \alpha$ (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Operationally, CP assigns each possible outcome a nonconformity score and retains those whose scores do not exceed a calibrated threshold q. Larger scores indicate poorer agreement with the fitted model. If R denotes the score of the realized outcome, then under a fixed distribution P, calibration seeks

$$\mathbb{P}_P\{R \leq q\} \geq 1 - \alpha. \tag{1}$$

In longitudinal care, however, a prediction set can affect the very outcome it is meant to cover. Consider vasopressor treatment in septic shock. At each decision point, a clinical model may predict the subsequent hemodynamic state under alternative treatments. If the set includes refractory hypotension at the current dose, the clinician may escalate support, and if it does not, escalation may be avoided. Either choice alters the patient's subsequent hemodynamic trajectory.

Because the conformal threshold q determines the prediction set, it also influences treatment and hence the deployment distribution. Schematically, $q \longrightarrow \pi_q \longrightarrow P_q$, where π_q is the treatment policy induced by threshold q, and P_q the resulting distribution of patient trajectories. The resulting distribution P_q in turn determines the coverage attained by q, feeding back into the criterion used to calibrate that same threshold. This creates a form of performative feedback and gives rise to a

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.

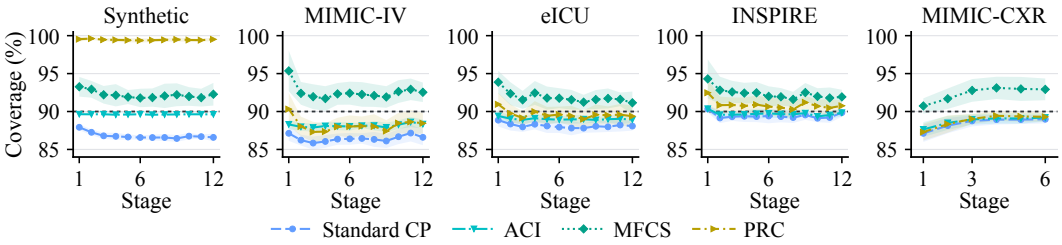

Figure 1: Stagewise coverage under sequential treatment feedback for Standard CP, ACI (Gibbs & Candes, 2021), MFCS (Prinster et al., 2024), and PRC (Li et al., 2025). Once prediction sets influence treatment, these methods exhibit varying degrees of coverage loss across stages.

self-consistent calibration problem. In contrast to Eq. 1, the relevant requirement is

$$\mathbb{P}_{P_q}\{R \leq q\} \geq 1 - \alpha. \tag{2}$$

Put differently, the threshold is not evaluated under a fixed deployment distribution, but rather helps determine that distribution through the treatment policy.

Our experiments give a first view of this effect, as shown in Figure 1. Naively ignoring the feedback can lead to substantial stagewise coverage errors, undermining the prediction reliability in deployment. In particular, standard CP can fall substantially below the nominal level once its prediction sets begin to influence treatment. A natural response to coverage loss is to recalibrate using newly observed outcomes (Gibbs & Candes, 2021; Li et al., 2025; Xu et al., 2024). In sequential care, however, these outcomes may become available only after further treatment decisions have been made, delaying correction of coverage errors. Alternatively, historical trajectories can be used to calibrate for a specified treatment policy (Tibshirani et al., 2019; Zhang et al., 2023). This permits calibration before deployment, but the resulting guarantee is tied to that policy. Under performative feedback, however, the target policy is not fixed independently of calibration. A threshold calibrated for one policy need not retain its coverage under the policy induced by its own use.

This naturally raises two questions. *(i) Can coverage under a treatment process induced by prediction sets be recovered from historical trajectories before deployment? (ii) Can a predictive system that changes treatment remain reliable under the very distribution it helps create?*

The first is an identification problem. Historical trajectories reflect the treatments that were actually given, whereas deployment may follow a different treatment process once prediction sets enter the decision. The sequential structure nevertheless provides a bridge, since coverage at stage t depends only on treatment decisions made up to that stage. We show that correcting for exactly these decisions recovers the corresponding deployment coverage from historical trajectories.

And the second is a calibration problem. Recall from Eq. 1 that, under a fixed distribution, the coverage map $q \mapsto \mathbb{P}_P\{R \leq q\}$ is monotone, so calibration reduces to quantile inversion. Here, the relevant map $q \mapsto \mathbb{P}_{P_q}\{R \leq q\}$ need not be monotone. Ordinary quantile calibration therefore no longer applies. Instead, the threshold must satisfy the coverage target under the distribution associated with that same threshold. This is the self-consistency at the heart of our problem.

1.1 Our contribution

With this context in mind, we propose SCCP, short for *Self-Consistent Conformal Prediction*, a calibration scheme that effectivety handles the performative feedback in sequential care. For intuition, we view the feedback loop as a matrix of coverage values, and its rows index the threshold governing treatment while its columns the cutoff used to assess coverage. The central idea of SCCP is to study this relation off the diagonal, where the two roles are separated, and then return to the diagonal, where they coincide. This construction yields a stage-specific threshold schedule and permits calibration before deployment. As a result, SCCP yields a calibration procedure that accounts for the distributional consequences of its own use before deployment.

Our contributions are threefold. First, we formulate conformal prediction under sequential treatment as a self-consistent calibration problem. We show that the coverage of any candidate threshold, in-

cluding the effect of preceding treatment decisions induced by earlier thresholds, is identifiable from historically logged trajectories. The formulation also makes clear why naivety increasing a threshold need not improve deployment coverage, since changing the threshold also changes the induced treatment distribution. Second, we propose SCCP, a calibration method accounting for the distributional consequences of its own use. We prove asymptotic marginal coverage at each treatment stage without requiring post-deployment outcomes. Third, we conduct a comprehensive evaluation in a synthetic benchmark, and semi-synthetic treatment settings built from four real-world clinical datasets. Across different intensities of feedback, SCCP maintains coverage close to the nominal level while retaining informative prediction sets. These results demonstrate the potential of self-consistent calibration for reliable uncertainty quantification in sequential clinical decision making. Code and supplementary materials are included in the anonymous supplementary submission.

2 PRELIMINARIES

2.1 RECAP: STANDARD CONFORMAL PREDICTION

We start by reviewing the core concepts of split-conformal prediction. Let $\mathcal{D}_{\mathrm{tra}}$ and $\mathcal{D}_{\mathrm{cal}}$ denote a proper training set and an independent calibration set, with $m = |\mathcal{D}_{\mathrm{cal}}|$. A predictive model is fit to training set and then held fixed. Given a nonconformity score $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, define $R_i = r(X_i, Y_i)$ for each calibration data (X_i, Y_i). Write $R_{(1)} \leq \cdots \leq R_{(m)}$ denote the ordered calibration scores. For a target miscoverage level $\alpha \in (0, 1)$, split CP takes $\widehat{q}_{1-\alpha} = R_{(\lceil (m+1)(1-\alpha) \rceil)}$, with the convention $R_{(m+1)} = +\infty$, and returns prediciton set, $\widehat{C}_\alpha(x) = \{y \in \mathcal{Y} : r(x, y) \leq \widehat{q}_{1-\alpha}\}$. If the calibration data and a future observation are *exchangeable*, then we have

$$\mathbb{P}\left\{Y_{m+1} \in \widehat{C}_\alpha(X_{m+1})\right\} \geq 1 - \alpha. \tag{3}$$

2.2 CONFORMAL PREDICTION FOR SEQUENTIAL CARE

We now consider conformal prediction for sequential care. A patient trajectory under a T-stage treatment process is written as $\tau = \big(S_0, A_0, Y_1, S_1, \ldots, S_{T-1}, A_{T-1}, Y_T, S_T\big)$, where $S_t \in \mathcal{S}$ contains the information available before treatment at stage t, $A_t \in \mathcal{A}$ is the selected treatment, and $Y_{t+1} \in \mathcal{Y}$ is the subsequent outcome. At each stage, (S_t, A_t) plays the role of the conformal input and Y_{t+1} that of the response. We write $R_t = r(S_t, A_t, Y_{t+1})$ for the nonconformity score and define the prediction set

$$C_q(s, a) = \{y \in \mathcal{Y} : r(s, a, y) \leq q\}. \tag{4}$$

Suppose the treatment policy π is fixed independently of calibration, and both calibration and future patients follow the induced trajectory distribution P_π. Split CP can then be applied across patients at each stage. If $\widehat{q}_t^{\pi}$ denotes the resulting threshold, then $\mathbb{P}_{\tau \sim P_\pi}\left\{Y_{t+1} \in C_{\widehat{q}_t^{\pi}}(S_t, A_t)\right\} \geq 1 - \alpha$. Thus, sequential dependence within an individual trajectory does not by itself invalidate stagewise conformal calibration. The key requirement is that calibration and deployment patients are generated under the same induced treatment process.

When prediction sets affect treatment process. The previous setting assumes that the treatment policy is fixed independently of the prediction sets. We now consider the setting where the prediction set itself influences treatment decisions, creating a performative feedback loop in sequential care. We observe N independent patient trajectories collected under a historical treatment policy $\mu_t(a \mid s) = \mathbb{P}(A_t = a \mid S_t = s)$. Writing K_t for the conditional distribution of the subsequent outcome and state, the historical trajectory distribution is

$$P_\mu(\tau) = p_0(s_0) \prod_{t=0}^{T-1} \mu_t(a_t \mid s_t) K_t(y_{t+1}, s_{t+1} \mid s_t, a_t).$$

Then consider a family of treatment policies $\{\pi_{t,q} : q \geq 0\}$ specified before calibration, where $\pi_{t,q}(a \mid s)$ is the probability of treatment a when threshold q is used at stage t. For a threshold schedule $\boldsymbol{q} = (q_0, \ldots, q_{T-1})$, the resulting trajectory distribution is

$$P_{\boldsymbol{q}}(\tau) = p_0(s_0) \prod_{t=0}^{T-1} \pi_{t,q_t}(a_t \mid s_t) K_t(y_{t+1}, s_{t+1} \mid s_t, a_t).$$

The two trajectory distributions are generally different because they induce different treatment assignment probabilities. Since only trajectories from P_μ are available before deployment, the challenge is to recover coverage under the threshold-induced distribution $P_{\boldsymbol{q}}$ from historical trajectories.

Coverage Objective. When the treatment policy is induced by the prediction set, the threshold used for calibration is no longer independent of the distribution under which coverage is evaluated. The conformal threshold therefore affects both the prediction set and the treatment process that determines future outcomes. Let $P_t(\boldsymbol{q}_{0:t})$ denote the induced distribution of (S_t, A_t, Y_{t+1}) after applying the threshold sequence up to stage t. Our goal is to construct $\boldsymbol{q}$ such that

$$\mathrm{Cov}_t(\boldsymbol{q}_{0:t}) := \mathbb{P}_{P_t(\boldsymbol{q}_{0:t})}\{Y_{t+1} \in C_{q_t}(S_t, A_t)\} \geq 1-\alpha, \qquad t = 0, \ldots, T-1. \tag{5}$$

During deployment, the stage-t predictive output is the corresponding set $C_{\widehat{q}_t}(S_t, A_t)$. Unlike standard CP, the evaluation distribution in this probability depends on the same threshold being calibrated. This coupling leads to a self-consistent calibration problem.

3 How existing methods handle the decision feedback

Having stated the self-consistent coverage objective above, we now place it in the context of existing approaches for handling distribution changes induced by prediction and decision making.

A broad class of conformal methods can be viewed as finding a threshold $q^\star$ satisfying

$$\mathbb{P}_{P^\star}\{R \leq q^\star\} \geq 1-\alpha, \tag{6}$$

where the main difference across methods lies in how the evaluation distribution $P^\star$ is obtained. Existing methods can largely be grouped into variants of adaptive conformal prediction, off-policy conformal prediction, and performative calibration (see Appendix F for further discussion).

(i) Adaptive conformal prediction. Adaptive conformal methods address evolving data distributions by updating the calibration rule as new observations become available. A generic form is $\widehat{q}^{\,\mathrm{adapt}}_{k,t} = Q_t(\mathcal{D}_{\mathrm{cal}}, \mathcal{H}_{k-1})$, where $\mathcal{H}_{k-1}$ contains observations revealed before update k.

For example, adaptive conformal inference (ACI) (Gibbs & Candes, 2021) adjusts the nominal miscoverage level in response to past coverage errors. Sequential predictive conformal methods instead exploit temporal dependence in past residuals to estimate future score quantiles, with MultiDimSPCI (Xu et al., 2024) extending this idea to multivariate outcomes through ellipsoidal prediction regions. These methods handle changing distributions by using outcomes observed after deployment. In contrast, our goal is to calibrate before deployment when such feedback is unavailable.

(ii) Off-policy conformal prediction. Off-policy conformal methods address a different type of distribution shift: calibrating predictions under a target policy using data collected under a historical policy. Their target distribution takes the form $P^\star = P_\pi$, where the target policy π is specified independently of calibration. Weighted conformal prediction transports calibration using density ratios between the target and historical distributions (Tibshirani et al., 2019). Conformal off-policy prediction constructs prediction intervals for the return of a specified target policy using off-policy data (Zhang et al., 2023). These approaches allow calibration under a different treatment process, but the target policy is fixed before calibration.

(iii) Performative conformal calibration. Performative methods consider settings where the object being calibrated changes the distribution under evaluation. Their target distribution can therefore be written as $P^\star = P_\theta$, where θ denotes the deployed decision rule. Performative risk control studies analogous risk constraints through repeated deployment and recalibration (Li et al., 2025). Conformal policy control considers policy adaptation relative to a safe reference policy (Prinster et al., 2026), while recent work on policy-coupled conformal prediction studies coverage of the realized outcome under actions induced by the prediction sets themselves (Zheng & Jin, 2026).

These works establish important connections between conformal calibration and decision feedback. Our setting adds a longitudinal aspect to this coupling. At the first stage, calibration reduces to a one-stage policy-coupled problem. For $t \geq 1$, however, earlier thresholds affect treatment decisions

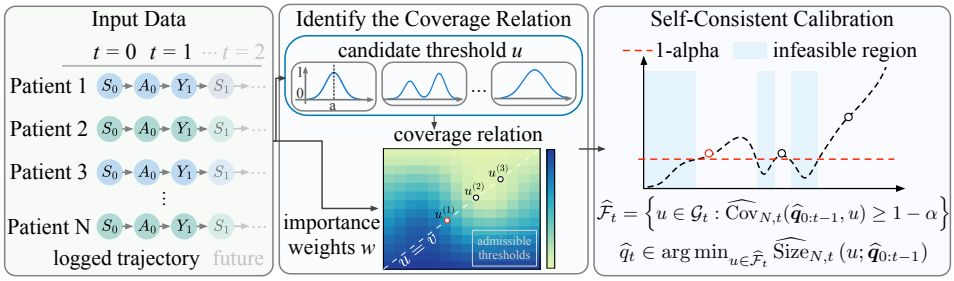

Figure 2: Illustration of SCCP. Identifying the coverage relation from historical trajectories and calibrating self-consistent thresholds in treatment order.

and thereby the state distribution on which later prediction sets are evaluated. Thus, the stage-t deployment distribution depends on the entire threshold prefix $\boldsymbol{q}_{0:t}$, rather than on q_t alone. Our contribution is to identify the corresponding stagewise coverage from historical trajectories and recursively calibrate the threshold schedule without observing outcomes.

4 Method

Recall the target in Eq. 5. The key challenge is that the target distribution $P_t(\boldsymbol{q}_{0:t})$ is not observed during calibration; the available trajectories instead follow P_μ. Our method, coined SCCP, addresses this in two steps (illustrated in Figure 2). We first recover the coverage relation $\mathcal{C}_t(u, v; \boldsymbol{q}_{0:t-1})$ from historical trajectories, keeping the treatment threshold u and the coverage cutoff v separate. We then set $u = v$ and calibrate the thresholds in treatment order.

4.1 Identifying the Coverage Relation

Fix the preceding thresholds $\boldsymbol{q}_{0:t-1}$ and consider using threshold u at stage t. We wish to evaluate the resulting coverage, although the available trajectories were generated under the historical policy μ. To separate the threshold's role in treatment from its role in coverage evaluation, let v denote the cutoff used to assess coverage. We define the coverage relation,

$$\mathcal{C}_t\left(u, v; \boldsymbol{q}_{0:t-1}\right) := \mathbb{P}_{P_t(\boldsymbol{q}_{0:t-1}, u)}\left\{R_t \le v\right\}. \tag{7}$$

For $t = 0$, the preceding threshold sequence is empty.

We now show that this coverage relation can be identified from historical trajectories. The key observation is that changing the policy changes the treatment probabilities, while the conditional law of future outcomes given the current state and treatment remains unchanged. We can therefore reweight historical trajectories by

$$W_t(u; \boldsymbol{q}_{0:t-1}) = \left\{\prod_{h=0}^{t-1} \frac{\pi_{h,q_h}(A_h \mid S_h)}{\mu_h(A_h \mid S_h)}\right\} \frac{\pi_{t,u}(A_t \mid S_t)}{\mu_t(A_t \mid S_t)}, \tag{8}$$

where the empty product equals one for $t = 0$. The preceding factors account for treatment decisions leading to the current state, while the final factor accounts for the treatment preceding the next outcome. Thus, deployment coverage can be expressed in terms of historical trajectories.

We require an overlap condition: any treatment that may be chosen by a candidate policy has positive probability under the historical policy at the relevant state. The following proposition shows that the desired coverage can then be recovered from historical trajectories, with proof in Appendix A.1.

Proposition 1 (Coverage identification)**.** *Coverage under an admissible treatment policy can be identified from the historical law P_μ. Specifically, for any admissible preceding thresholds $\boldsymbol{q}_{0:t-1}$, treatment threshold u, and coverage cutoff v,*

$$\mathcal{C}_t\left(u, v; \boldsymbol{q}_{0:t-1}\right) = \mathbb{E}_{P_\mu}\left[W_t\left(u; \boldsymbol{q}_{0:t-1}\right) \mathbf{1}\{R_t \le v\}\right]. \tag{9}$$

The proposition is the bridge between deployment and historical data. The left-hand side is coverage under a treatment process that has not yet been observed, while the right-hand side depends only on

trajectories drawn from P_μ. Thus, for any treatment threshold u, we can evaluate the corresponding deployment coverage before that threshold is actually used.

Empirical estimation. In practice, only a finite calibration cohort is available. We therefore estimate the identified coverage relation by a weighted empirical average over historical trajectories. Suppose that N independent calibration trajectories $\{\tau_i\}_{i=1}^N$ are available. If the historical policy is unknown, we estimate it from data separate from the final calibration sample and denote the estimate by $\widehat{\mu}$. Let $\widehat{W}_{i,t}$ be the corresponding weight for trajectory i. We estimate the coverage relation as

$$\widehat{\mathcal{C}}_{N,t}(u, v; \boldsymbol{q}_{0:t-1}) = \frac{\sum_{i=1}^N \widehat{W}_{i,t}(u; \boldsymbol{q}_{0:t-1}) \mathbf{1}\{R_{i,t} \le v\}}{\sum_{i=1}^N \widehat{W}_{i,t}(u; \boldsymbol{q}_{0:t-1})}. \tag{10}$$

In other words, this estimator reweights historical trajectories to estimate the coverage that would be obtained if threshold u were used to guide treatment.

4.2 Self-consistent Calibration

So far, u and v have been separated to identify the coverage relation. In deployment, however, there is only one threshold: it both influences treatment and defines the prediction set. We therefore evaluate the diagonal of the coverage relation, where the treatment threshold and coverage cutoff coincide, and write

$$\widehat{\mathrm{Cov}}_{N,t}(\boldsymbol{q}_{0:t-1}, u) := \widehat{\mathcal{C}}_{N,t}(u, u; \boldsymbol{q}_{0:t-1}). \tag{11}$$

For a candidate threshold u, this quantity asks whether using u at stage t would achieve the desired coverage under the treatment process induced by that same threshold.

At stage t, the preceding thresholds $\widehat{\boldsymbol{q}}_{0:t-1}$ are already fixed. We determine the current threshold by imposing the self-consistent coverage constraint. Let $\mathcal{G}_t$ denote the set of admissible thresholds considered at this stage. The feasible set is defined as

$$\widehat{\mathcal{F}}_t = \left\{ u \in \mathcal{G}_t : \widehat{\mathrm{Cov}}_{N,t}(\widehat{\boldsymbol{q}}_{0:t-1}, u) \ge 1 - \alpha \right\}. \tag{12}$$

Among the feasible thresholds, we prefer the most informative prediction sets. Let

$$\widehat{\mathrm{Size}}_{N,t}(u; \boldsymbol{q}_{0:t-1}) = \frac{\sum_{i=1}^N \widehat{W}_{i,t}(u; \boldsymbol{q}_{0:t-1}) \,\mathrm{size}\big(C_u(S_{i,t}, A_{i,t})\big)}{\sum_{i=1}^N \widehat{W}_{i,t}(u; \boldsymbol{q}_{0:t-1})}. \tag{13}$$

We then select $\widehat{q}_t \in \arg\min_{u \in \widehat{\mathcal{F}}_t} \widehat{\mathrm{Size}}_{N,t}(u; \widehat{\boldsymbol{q}}_{0:t-1})$.

The thresholds are calibrated in treatment order. At stage 0, there are no preceding thresholds, so we first choose $\widehat{q}_0$. Once it is fixed, the treatment decisions induced by $\widehat{q}_0$ affect the states encountered at stage 1. We therefore calibrate $\widehat{q}_1$ under this new state distribution. Repeating this procedure yields the recursive construction: $\widehat{q}_0 \longrightarrow \widehat{q}_1 \longrightarrow \cdots \longrightarrow \widehat{q}_{T-1}$. Because stage-$t$ coverage depends only on thresholds selected up to stage t, later threshold choices cannot change coverage constraints already enforced at earlier stages. The resulting schedule is $\widehat{\boldsymbol{q}} = (\widehat{q}_0, \ldots, \widehat{q}_{T-1})$, which defines the deployed prediction set $C_{\widehat{q}_t}(S_t, A_t)$ at each stage. Under a fixed distribution, increasing the conformal threshold enlarges the prediction set and cannot reduce coverage. Here, however, the threshold also changes treatment and hence the score distribution. Consequently, the self-consistent coverage may be nonmonotone in u (as further discussed in Appendix B). For this reason, ordinary quantile inversion is no longer applicable, so the coverage constraint must be evaluated directly.

Coverage guarantee. One final issue remains, namely whether a threshold satisfying the empirical coverage constraint also satisfies the corresponding population constraint. Let $\mathcal{G}_{0:t} = \mathcal{G}_0 \times \cdots \times \mathcal{G}_t$ and define the largest coverage estimation error by

$$\Delta_N = \max_{0 \le t < T} \max_{\boldsymbol{q}_{0:t} \in \mathcal{G}_{0:t}} \left| \widehat{\mathrm{Cov}}_{N,t}(\boldsymbol{q}_{0:t-1}, q_t) - \mathrm{Cov}_t(\boldsymbol{q}_{0:t}) \right|.$$

Proposition 2 (Stagewise marginal coverage)**.** *For fixed T and finite candidate grids, assume overlap, bounded cumulative likelihood ratios, and consistent estimation of the historical policy. If the procedure returns a complete threshold schedule with probability tending to one, then*

$$\min_{0 \le t < T} \mathrm{Cov}_t\left(\widehat{\boldsymbol{q}}_{0:t}\right) \ge 1 - \alpha - \Delta_N.$$

Moreover, $\Delta_N = o_{\mathbb{P}}(1)$ *as the calibration sample grows, and hence*

$$\min_{0 \le t < T} \mathrm{Cov}_t\left(\widehat{\boldsymbol{q}}_{0:t}\right) \ge 1 - \alpha - o_{\mathbb{P}}(1). \tag{14}$$

In other words, as the calibration sample grows, the coverage shortfall of the calibrated schedule vanishes at every treatment stage. The proof is given in Appendix A.3.

5 Numerical Studies

5.1 Setup

We conduct a series of experiments to evaluate the reliability and efficiency of SCCP under sequential treatment feedback. The experimental setup is described below.

Benchmarks. We evaluate SCCP on one synthetic benchmark and four semi-synthetic clinical benchmarks. Each is formulated as a sequential treatment problem with a state S_t, a categorical action A_t, and a two-dimensional post-treatment response Y_{t+1}. (i) Synthetic. The benchmark simulates sequential management of a critically ill patient, where treatment decisions based on the current state influence future disease severity and prediction errors. It contains 12 treatment stages, three possible actions, and two correlated outcomes, with feedback strength and direction varied. (ii) MIMIC-IV (Johnson et al., 2024; 2023). Each trajectory contained 12 consecutive 4-hour intervals. The responses are hypotension and tachycardia burdens, and actions represent joint fluid and vasopressor treatment intensity. (iii) eICU (Pollard et al., 2018). We construct the same 12-stage, 4-hour sequential treatment task from eICU records. (iv) INSPIRE (Lim et al., 2024; Lim & Lee, 2026). Each perioperative trajectory consists of 12 consecutive 10-minute intervals. Responses are hypotension and hypertension burdens, and actions correspond to no treatment, fluid only, or vasopressor-containing treatment. (v) MIMIC-CXR (Johnson et al., 2019a;b). Each trajectory consists of 6 consecutive 6-hour intervals. Responses are hypoxaemia and tachypnoea burdens, and actions correspond to no respiratory support, non-invasive support, or invasive ventilation. Dataset construction and treatment definitions are detailed in Appendix D.2.

Baselines. We use Standard CP as the primary feedback-agnostic baseline: it uses the same historical calibration trajectories as SCCP but ignores threshold-induced treatment feedback. We also compare with ACI (Gibbs & Candes, 2021), MFCS (Prinster et al., 2024), and PRC (Li et al., 2025), representing online adaptation, multistep feedback-shift correction, and performative risk calibration, respectively. For PRC, we use a stage-wise adaptation to match our stagewise marginal coverage target. Information requirements and adaptations are detailed in Appendix D.2.

Metrics. We use the scale-normalized score (Romano et al., 2019), $R_t = \max_{j\in\{1,2\}} |Y_{t+1,j} - \widehat{m}_j(S_t, A_t)|/\widehat{\sigma}_j(S_t, A_t)$, with target coverage $1 - \alpha = 0.90$. For run s, let $\widehat{C}_{s,t}$ denote empirical coverage at stage t. We define worst-stage coverage $\mathrm{WSC}_s = \min_t \widehat{C}_{s,t}$ and mean coverage $\mathrm{MeanCov}_s = T^{-1}\sum_t \widehat{C}_{s,t}$. Efficiency is measured by normalized width, $\mathrm{NWidth}_{s,i,t} = \frac{1}{2}\sum_{j=1}^{2} 2\widehat{q}_{s,i,t}\widehat{\sigma}_j(S_{s,i,t}, A_{s,i,t})/\widehat{\sigma}^{\mathrm{out}}_{s,j}$, where $\widehat{\sigma}^{\mathrm{out}}_{s,j}$ is the sample standard deviation of outcome j in the prediction-model split and $\widehat{q}_{s,i,t}$ is the threshold applied to patient i at stage t. For methods returning a fixed stagewise schedule, $\widehat{q}_{s,i,t} = \widehat{q}_{s,t}$. We report WSC, MeanCov, and NWidth as mean $\pm$ standard deviation across runs; smaller NWidth indicates more informative prediction sets. Statistical details are given in Appendix D.2.

5.2 Results

Overall performance. In Table 1, together with Figures 3 (and Appendix), we observe that SCCP consistently maintains coverage close to the nominal level while producing compact prediction sets. Standard CP and ACI often return narrower sets, but their worst-stage coverage remains below the target, indicating that this apparent gain in efficiency comes at the cost of unreliable coverage. MFCS restores coverage more conservatively and requires noticeably wider prediction sets. PRC can achieve high coverage in some settings through repeated deployment updates, but this requires post-deployment outcomes and can lead to conservative prediction sets. Among methods whose coverage reaches the nominal 90% target, SCCP achieves the smallest normalized width across

Table 1: Main results across all methods. Values are reported as mean ± standard deviation. Green marks coverage above 90%; **boldface** marks the most efficient WSC-qualified method.

Metric	Method	Synthetic	MIMIC-IV	eICU	INSPIRE	MIMIC-CXR
MeanCov (%)	Standard CP	86.8 ± 0.1	86.4 ± 0.5	88.1 ± 0.4	89.4 ± 0.2	88.5 ± 0.8
	ACI	89.9 ± 0.0	88.1 ± 0.2	89.0 ± 0.2	89.7 ± 0.2	88.7 ± 0.7
	MFCS	92.2 ± 0.9	92.5 ± 1.2	91.9 ± 0.9	92.4 ± 1.0	92.4 ± 1.4
	PRC	99.5 ± 0.0	88.2 ± 0.6	89.6 ± 0.6	90.9 ± 0.5	88.8 ± 0.8
	SCCP	**90.4 ± 0.3**	**90.3 ± 0.2**	**90.3 ± 0.2**	**90.3 ± 0.3**	**90.3 ± 0.4**
WSC (%)	Standard CP	86.4 ± 0.4	85.8 ± 0.9	87.8 ± 0.6	89.1 ± 0.5	87.1 ± 1.1
	ACI	89.9 ± 0.1	87.9 ± 0.5	88.9 ± 0.4	89.4 ± 0.4	87.6 ± 1.0
	MFCS	91.8 ± 1.2	91.7 ± 1.6	91.2 ± 1.5	91.6 ± 1.0	90.7 ± 1.0
	PRC	99.4 ± 0.3	*87.3 ± 1.3*	89.0 ± 1.2	90.3 ± 1.1	*87.3 ± 1.1*
	SCCP	**90.0 ± 0.2**	**90.2 ± 0.2**	**90.2 ± 0.7**	**90.0 ± 0.2**	**90.2 ± 0.6**
NWidth	Standard CP	2.58 ± 0.06	4.09 ± 0.19	5.27 ± 0.43	6.55 ± 0.22	6.48 ± 1.23
	ACI	3.90 ± 0.08	4.46 ± 0.23	5.49 ± 0.47	6.63 ± 0.22	6.61 ± 1.26
	MFCS	4.49 ± 0.34	5.85 ± 0.66	6.48 ± 0.63	7.67 ± 0.50	8.42 ± 2.08
	PRC	9.71 ± 0.14	4.47 ± 0.22	5.66 ± 0.45	7.04 ± 0.29	6.59 ± 1.25
	SCCP	**3.84 ± 0.11**	**5.00 ± 0.31**	**5.87 ± 0.51**	**6.84 ± 0.23**	**7.34 ± 1.46**

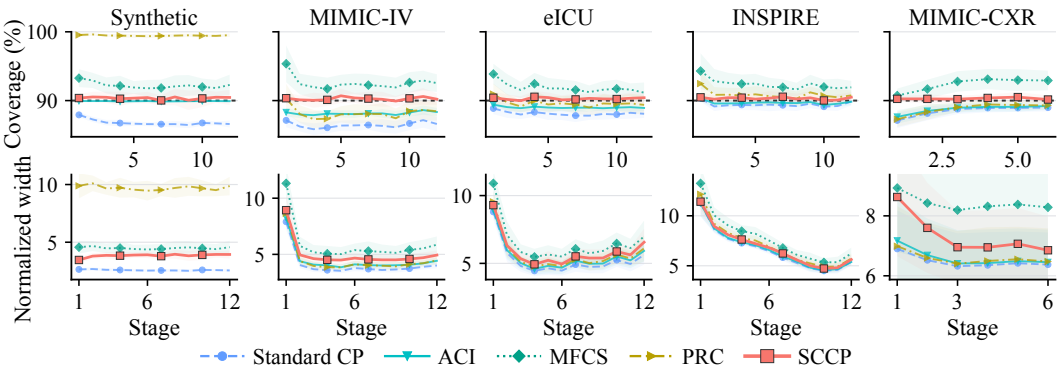

Figure 3: Stagewise reliability under sequential treatment feedback.

all five benchmarks. The stagewise results further show that this reliability persists throughout treatment, with no systematic deterioration at later stages.

Likewise, Figure 4 studies varying feedback strengths. SCCP remains close to the target coverage across different feedback regimes while adapting prediction-set width to the resulting uncertainty. In contrast, Standard CP and ACI are more sensitive to feedback because they do not calibrate the distribution induced by the prediction threshold. MFCS remains conservative, whereas PRC adapts through observed deployment outcomes.

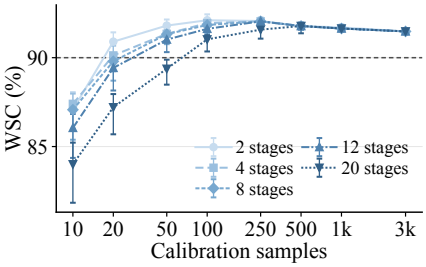

Figure 6: Calibration-size sensitivity under increasing horizon.

Ablation analysis. Figure 6 (additional datasets and baselines in Appendix E.1) shows that limited calibration data have a larger impact as the treatment horizon grows, while the gap diminishes with more calibration trajectories. This reflects the increasing difficulty of estimating later stage treatment corrections when the available calibration sample is limited. More importantly, the ablation results in Figure 5 (Appendix E.2) support the design of SCCP. Removing the current treatment correction (No-current) reduces worst-stage coverage, confirming that threshold-induced treatment must be accounted for. Retaining only the current correction (Current-only) ignores the state distribution shaped by earlier treatments. The Frozen variant further shows that calibrating under a fixed treatment process overlooks the coupling between calibration and treatment, leading to

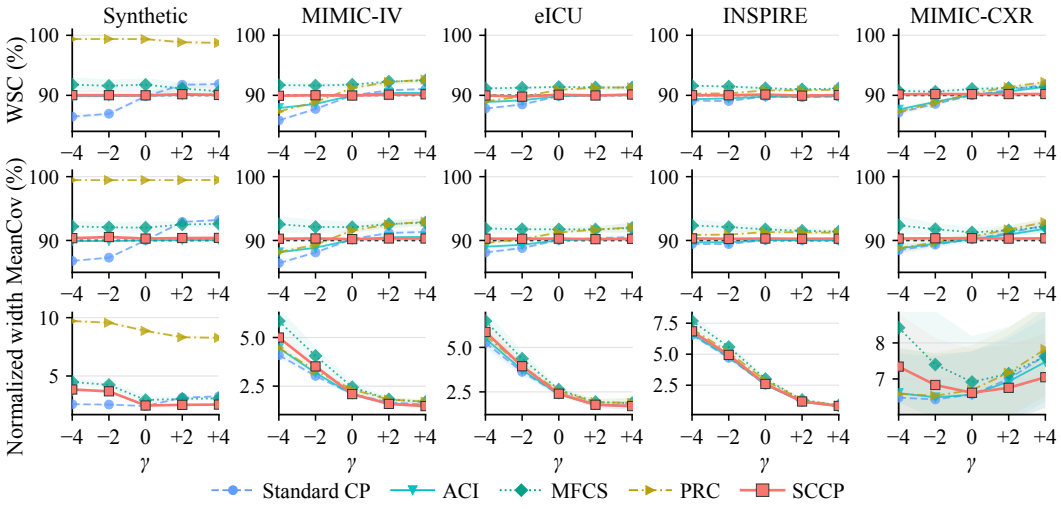

Figure 4: Sensitivity to signed feedback strength.

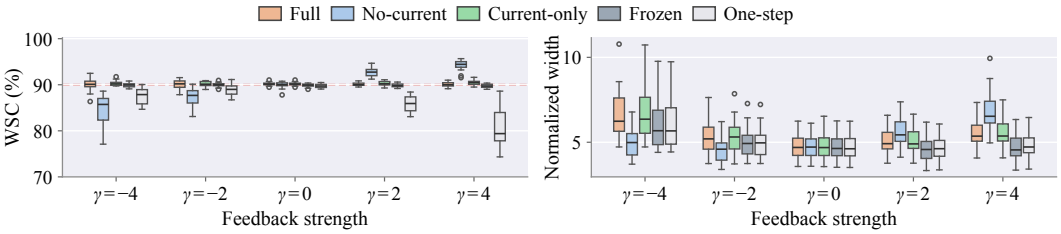

Figure 5: Ablation study on MIMIC-CXR.

coverage loss. Finally, the one-step correction underperforms the full method, highlighting the need for recursive calibration across stages. Together, these findings support incorporating the complete treatment history through the current stage for self-consistent calibration.

These findings highlight the importance of accounting for performative feedback in calibration. Together, they show that SCCP achieves a favorable balance between reliability and efficiency, providing a principled calibration scheme under performative feedback.

6 Discussion & Future Work

In this work, we studied conformal prediction for sequential treatment when prediction sets can influence the treatment process under which their coverage is evaluated. This leads to a self-consistent calibration problem in which the conformal threshold affects both the prediction set and the distribution of future outcomes. We proposed SCCP, which recovers stagewise coverage from historical trajectories and uses it to calibrate a threshold schedule before deployment. We proved asymptotic marginal coverage at each treatment stage under the resulting treatment process. Across extensive experiments, SCCP maintained coverage close to the nominal level while producing informative prediction sets. The results further showed the importance of accounting for both current and preceding treatment decisions under feedback. More broadly, our findings suggest that conformal uncertainty can be calibrated before deployment even when its use changes subsequent decisions, without requiring outcomes from the induced treatment process.

Future work includes extending the framework to richer and more personalized treatment policies, as well as prospective evaluation with clinical decision makers. We hope this work provides a useful foundation for uncertainty quantification in settings where predictions do not merely describe future outcomes, but also influence the decisions that shape them.

AI USE STATEMENT

In this work, we used ChatGPT 5.6 Sol for language editing and polishing, including improvements to grammar, clarity, and readability. We have not used it to develop research ideas or methodology, experiments, analyses, or conclusions. We reviewed all AI-assisted edits and verified that they preserve the intended technical content and do not introduce unsupported statements or incorrect citations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise specific ethical concerns.

REPRODUCIBILITY STATEMENT

We provide the implementation of SCCP and the experimental code in the anonymous supplementary materials. Detailed experimental settings, dataset construction and preprocessing procedures, and hyperparameter choices are described in the appendix and supplementary materials. Proofs of the theoretical results and the assumptions used in our analysis are also included in the appendix.

REFERENCES

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

Ilia Azizi, Juraj Bodik, Jakob M. Heiss, and Bin Yu. Clear: Calibrated learning for epistemic and aleatoric risk. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 157741–157813, 2026.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 6045–6061, 2022.

Santiago Cortes-Gomez, Carlos Patiño, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 4227–4246, 2025.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Konstantin Hess and Stefan Feuerriegel. Stabilized neural prediction of potential outcomes in continuous time. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 2131–2160, 2025.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR Database. *PhysioNet*, September 2019a. doi: 10.13026/C2JT1Q. URL https://doi.org/10.13026/C2JT1Q. Version 2.0.0.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019b.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10 (1):1, 2023.

Victor Li, Baiting Chen, Yuzhen Mao, Qi Lei, and Zhun Deng. Performative risk control: Calibrating models for reliable deployment under performativity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 145912–145943, 2025.

Leerang Lim and Hyung-Chul Lee. INSPIRE, a publicly available research dataset for perioperative medicine. *PhysioNet*, June 2026. doi: 10.13026/1eay-yc85. URL `https://doi.org/10.13026/1eay-yc85`. Version 1.4.2.

Leerang Lim, Hyeonhoon Lee, Chul-Woo Jung, Dayeon Sim, Xavier Borrat, Tom J. Pollard, Leo A. Celi, Roger G. Mark, Simon T. Vistisen, and Hyung-Chul Lee. INSPIRE, a publicly available research dataset for perioperative medicine. *Scientific Data*, 11:655, 2024.

Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: A novel orthogonal learner. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 105039–105089, 2024.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4929–4939, 2020.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pp. 7599–7609, 2020.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pp. 41086–41118, 2024.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Don Stanton. Conformal policy control. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, pp. 1–476. Springer Cham, 2005.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pp. 11559–11569, 2021.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pp. 38707–38727, 2023.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pp. 55076–55099, 2024.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 2751–2768, 2023.

Yurui Zheng and Ying Jin. Prediction sets for counterfactual decisions: Coverage, optimality, and conformal prediction, 2026. URL https://arxiv.org/abs/2607.02206.

APPENDIX

A MATHEMATICAL PROOFS

This section provides the proofs of the propostions in Section 4. Throughout, all models and policy components are treated as fixed after being fitted before the final calibration sample is used.

A.1 PROOF OF PROPOSITION 1

We first show how coverage under a candidate treatment process can be identified from trajectories collected under the historical policy, and then state the conditions required for this identification.

Assumption 1 (Longitudinal identification)**.** *The calibration trajectories are independent patient-level draws from the historical distribution P_μ. Changing the treatment policy does not change the conditional distribution of the next outcome and state given the current state and treatment:*

$$(Y_{t+1}, S_{t+1}) \mid (S_t, A_t) \sim K_t(\cdot \mid S_t, A_t).$$

We also assume overlap, whenever a candidate policy may assign positive probability to a treatment, that treatment also has positive probability under the historical policy,

$$\pi_{t,q}(a \mid s) > 0 \quad \Longrightarrow \quad \mu_t(a \mid s) > 0.$$

Proof. Fix a stage t, the preceding thresholds $\boldsymbol{q}_{0:t-1}$, a candidate threshold u for the current treatment, and a coverage cutoff v. Consider the trajectory prefix required to evaluate stage-t coverage, $\tau_{0:t+1} = (S_0, A_0, Y_1, S_1, \ldots, S_t, A_t, Y_{t+1}, S_{t+1})$. Under the historical policy, the probability of an observed trajectory prefix up to stage t can be written as

$$p_\mu(\tau_{0:t+1}) = p_0(s_0) \prod_{h=0}^{t} \mu_h(a_h \mid s_h) K_h(y_{h+1}, s_{h+1} \mid s_h, a_h).$$

If instead the thresholds $q_0, \ldots, q_{t-1}$ had been used at the previous stages and u were used at the current stage, the corresponding probability would be

$$\begin{aligned} p_{\boldsymbol{q}_{0:t-1},u}(\tau_{0:t+1}) = p_0(s_0) \prod_{h=0}^{t-1} \pi_{h,q_h}(a_h \mid s_h) K_h(y_{h+1}, s_{h+1} \mid s_h, a_h) \\ \times \pi_{t,u}(a_t \mid s_t) K_t(y_{t+1}, s_{t+1} \mid s_t, a_t). \end{aligned}$$

By Assumption 1, the initial-state distribution and the conditional transition terms are shared across the two treatment processes. Hence, on the support of the candidate process,

$$\begin{aligned} \frac{p_{\boldsymbol{q}_{0:t-1},u}(\tau_{0:t+1})}{p_\mu(\tau_{0:t+1})} &= \left\{ \prod_{h=0}^{t-1} \frac{\pi_{h,q_h}(A_h \mid S_h)}{\mu_h(A_h \mid S_h)} \right\} \frac{\pi_{t,u}(A_t \mid S_t)}{\mu_t(A_t \mid S_t)} \\ &= W_t(u; \boldsymbol{q}_{0:t-1}). \end{aligned}$$

The right-hand side is exactly the prefix importance weight $W_t(u; \boldsymbol{q}_{0:t-1})$ defined in Eq. 8.

Therefore, for any integrable function $f(S_t, A_t, Y_{t+1})$, the change-of-measure identity gives

$$\mathbb{E}_{P_t(\boldsymbol{q}_{0:t-1},u)}[f] = \mathbb{E}_{P_\mu}\left[W_t(u; \boldsymbol{q}_{0:t-1}) f\right]. \tag{A15}$$

Taking $f = \mathbf{1}\{R_t \le v\}$ yields

$$\begin{aligned} \mathcal{C}_t(u, v; \boldsymbol{q}_{0:t-1}) &= \mathbb{P}_{P_t(\boldsymbol{q}_{0:t-1},u)}\{R_t \le v\} \\ &= \mathbb{E}_{P_\mu}\left[W_t(u; \boldsymbol{q}_{0:t-1}) \mathbf{1}\{R_t \le v\}\right], \end{aligned}$$

Finally, taking $f \equiv 1$ in Eq. A15 gives

$$\mathbb{E}_{P_\mu}\left[W_t(u; \boldsymbol{q}_{0:t-1})\right] = 1.$$

□

A.2 Uniform Accuracy of the Coverage Estimator

The threshold schedule is selected using the calibration data. Therefore, it is not sufficient to show that the coverage estimator is accurate for one fixed threshold sequence. We need it to be accurate for every candidate sequence that the calibration procedure may select.

For brevity, write

$$W_t(\boldsymbol{q}_{0:t}) := W_t(q_t; \boldsymbol{q}_{0:t-1}), \qquad \widehat{W}_t(\boldsymbol{q}_{0:t}) := \widehat{W}_t(q_t; \boldsymbol{q}_{0:t-1}).$$

For each stage t, let $\mathcal{G}_t$ be the finite set of candidate thresholds considered by the procedure, and define $\mathcal{G}_{0:t} = \mathcal{G}_0 \times \cdots \times \mathcal{G}_t$.

We measure the largest coverage estimation error over all stages and all candidate prefixes by

$$\Delta_N := \max_{0 \le t < T} \max_{\boldsymbol{q}_{0:t} \in \mathcal{G}_{0:t}} \left| \widehat{\mathrm{Cov}}_{N,t}(\boldsymbol{q}_{0:t-1}, q_t) - \mathrm{Cov}_t(\boldsymbol{q}_{0:t}) \right|. \tag{A16}$$

Thus, Δ_N is the worst error made by the coverage estimator among all candidates that the procedure may consider.

Assumption 2 (Uniform estimation)**.** *The horizon T is fixed, and each candidate set $\mathcal{G}_t$ is finite and fixed before the final calibration sample is used. The cumulative importance weights are uniformly bounded: there exists a constant $B < \infty$ such that*

$$0 \le W_t(\boldsymbol{q}_{0:t}) \le B$$

for every stage and every candidate prefix.

If the historical policy is estimated, it is estimated using data independent of the final calibration trajectories. We further assume that the resulting estimated prefix weights satisfy

$$\max_{0 \le t < T} \max_{\boldsymbol{q}_{0:t} \in \mathcal{G}_{0:t}} \frac{1}{N} \sum_{i=1}^{N} \left| \widehat{W}_{i,t}(\boldsymbol{q}_{0:t}) - W_{i,t}(\boldsymbol{q}_{0:t}) \right| = o_{\mathbb{P}}(1). \tag{A17}$$

When the historical policy is known, this condition holds trivially because $\widehat{W}_{i,t} = W_{i,t}$.

This condition ensures that replacing μ with $\widehat{\mu}$ does not introduce asymptotically persistent error in the prefix weights. When μ is known, the condition holds trivially.

Lemma 1 (Uniform coverage accuracy)**.** *Under Assumptions 1 and 2,*

$$\Delta_N = o_{\mathbb{P}}(1).$$

Proof. We first consider the case in which the historical policy is known. For a candidate prefix $\boldsymbol{q}_{0:t}$, define

$$A_{N,t}(\boldsymbol{q}_{0:t}) = \frac{1}{N} \sum_{i=1}^{N} W_{i,t}(\boldsymbol{q}_{0:t}) \mathbf{1}\{R_{i,t} \le q_t\}, \qquad B_{N,t}(\boldsymbol{q}_{0:t}) := \frac{1}{N} \sum_{i=1}^{N} W_{i,t}(\boldsymbol{q}_{0:t}).$$

The corresponding empirical self-consistent coverage is

$$\widehat{\mathrm{Cov}}_{N,t}(\boldsymbol{q}_{0:t-1}, q_t) = \frac{A_{N,t}(\boldsymbol{q}_{0:t})}{B_{N,t}(\boldsymbol{q}_{0:t})}.$$

By Proposition 1, $\mathbb{E}_{P_\mu}\left[W_t(\boldsymbol{q}_{0:t}) \mathbf{1}\{R_t \le q_t\}\right] = \mathrm{Cov}_t(\boldsymbol{q}_{0:t})$, and $\mathbb{E}_{P_\mu}\left[W_t(\boldsymbol{q}_{0:t})\right] = 1$.

Hence, for every fixed candidate prefix, the law of large numbers gives

$$A_{N,t}(\boldsymbol{q}_{0:t}) \xrightarrow{\mathbb{P}} \mathrm{Cov}_t(\boldsymbol{q}_{0:t}), \qquad B_{N,t}(\boldsymbol{q}_{0:t}) \xrightarrow{\mathbb{P}} 1.$$

Because T and all candidate grids are fixed and finite, the total number of candidate prefixes is finite. The above convergence therefore holds simultaneously over all stages and candidate prefixes:

$$\max_{t,\boldsymbol{q}_{0:t}} |A_{N,t}(\boldsymbol{q}_{0:t}) - \mathrm{Cov}_t(\boldsymbol{q}_{0:t})| = o_{\mathbb{P}}(1),$$

and

$$\max_{t,\boldsymbol{q}_{0:t}} |B_{N,t}(\boldsymbol{q}_{0:t}) - 1| = o_{\mathbb{P}}(1).$$

In particular, the denominators are bounded away from zero with probability tending to one. It follows that

$$\max_{t,\boldsymbol{q}_{0:t}} \left| \widehat{\mathrm{Cov}}_{N,t}(\boldsymbol{q}_{0:t-1}, q_t) - \mathrm{Cov}_t(\boldsymbol{q}_{0:t}) \right| = o_{\mathbb{P}}(1).$$

If the historical policy is estimated, Assumption 2 implies that replacing $W_{i,t}$ by $\widehat{W}_{i,t}$ changes both the numerator and denominator by $o_{\mathbb{P}}(1)$ uniformly over the finite candidate set. The same conclusion therefore holds for the estimated-weight coverage estimator.

Thus, $\Delta_N = o_{\mathbb{P}}(1)$. □

A finite-sample bound. The preceding consistency result can also be made explicit when the historical policy is known. Let $M = \sum_{t=0}^{T-1} \prod_{h=0}^{t} |\mathcal{G}_h|$ denote the total number of candidate threshold prefixes, and suppose $0 \le W_t(\boldsymbol{q}_{0:t}) \le B$ uniformly as in Assumption 2. For $\delta \in (0,1)$, define $\varepsilon_N = B\sqrt{\frac{\log(4M/\delta)}{2N}}$. If $\varepsilon_N < 1$, then with probability at least $1-\delta$, $\Delta_N \le \frac{2\varepsilon_N}{1-\varepsilon_N}$. Consequently, whenever SCCP returns a complete threshold schedule,

$$\min_{0 \le t < T} \mathrm{Cov}_t(\widehat{\boldsymbol{q}}_{0:t}) \ge 1 - \alpha - \frac{2\varepsilon_N}{1-\varepsilon_N}$$

with probability at least $1-\delta$.

A.3 Proof of Proposition 2

Proof. Let $\mathcal{E}_N$ denote the event that SCCP returns a complete threshold schedule. On $\mathcal{E}_N$, every selected threshold satisfies the empirical coverage constraint,

$$\widehat{\mathrm{Cov}}_{N,t}(\widehat{\boldsymbol{q}}_{0:t-1}, \widehat{q}_t) \ge 1 - \alpha.$$

The selected prefix $\widehat{\boldsymbol{q}}_{0:t}$ depends on the calibration data. This does not cause a problem because Δ_N controls the estimation error simultaneously over all candidate prefixes. Hence,

$$\mathrm{Cov}_t(\widehat{\boldsymbol{q}}_{0:t}) \ge \widehat{\mathrm{Cov}}_{N,t}(\widehat{\boldsymbol{q}}_{0:t-1}, \widehat{q}_t) - \Delta_N.$$

Combining this inequality with empirical feasibility gives

$$\mathrm{Cov}_t(\widehat{\boldsymbol{q}}_{0:t}) \ge 1 - \alpha - \Delta_N$$

at every stage for which the complete schedule is returned. Hence,

$$\min_{0 \le t < T} \mathrm{Cov}_t(\widehat{\boldsymbol{q}}_{0:t}) \ge 1 - \alpha - \Delta_N. \tag{A18}$$

By Lemma 1, $\Delta_N = o_{\mathbb{P}}(1)$. Therefore, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\mathcal{E}_N \cap \left\{ \min_{0 \le t < T} \mathrm{Cov}_t(\widehat{\boldsymbol{q}}_{0:t}) \ge 1 - \alpha - \varepsilon \right\} \right) \longrightarrow 1.$$

Equivalently,

$$\min_{0 \le t < T} \mathrm{Cov}_t(\widehat{\boldsymbol{q}}_{0:t}) \ge 1 - \alpha - o_{\mathbb{P}}(1).$$

□

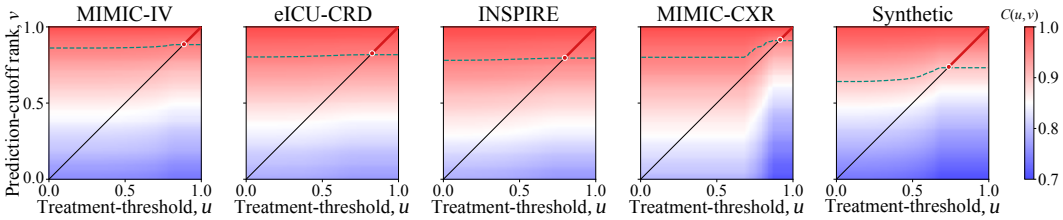

Figure A6: **Geometry of the coverage relation across five datasets.** Heatmaps show $\mathcal{C}_t(u, v)$ as the treatment threshold u and coverage cutoff v vary; the black diagonal corresponds to $u = v$. **Takeaway.** Coverage increases with v when treatment is fixed, whereas along $u = v$ this monotonic pattern can be altered by treatment feedback.

A.4 Remark: Availability of a Complete Schedule

The coverage guarantee above and the existence of a complete threshold schedule are distinct. Proposition 2 states that, whenever SCCP returns a threshold at every stage, the resulting schedule satisfies the desired stagewise coverage asymptotically. A feasible candidate may nevertheless fail to exist at some stage for a finite calibration sample.

A simple sufficient condition for asymptotic availability is that every prefix that can be reached by the calibration procedure admits at least one candidate with a fixed population coverage margin. Specifically, suppose that there exists $\eta > 0$ such that, for every reachable $\boldsymbol{q}_{0:t-1}$, some $u \in \mathcal{G}_t$ makes the extended prefix $\boldsymbol{q}_{0:t} = (\boldsymbol{q}_{0:t-1}, u)$ satisfy

$$\mathrm{Cov}_t(\boldsymbol{q}_{0:t}) \geq 1 - \alpha + \eta.$$

Whenever $\Delta_N < \eta$, this candidate remains empirically feasible because

$$\begin{aligned} \widehat{\mathrm{Cov}}_{N,t}(\boldsymbol{q}_{0:t-1}, u) &\geq \mathrm{Cov}_t(\boldsymbol{q}_{0:t}) - \Delta_N \\ &\geq 1 - \alpha. \end{aligned}$$

Thus, on the event $\{\Delta_N < \eta\}$, a feasible candidate exists at every reachable stage. Applying this argument recursively along the selected threshold sequence yields a complete schedule. Since $\Delta_N = o_{\mathbb{P}}(1)$,

$$\mathbb{P}(\mathcal{E}_N) \longrightarrow 1.$$

In our experiments, we record whether a complete schedule is obtained rather than discarding unavailable runs. Empirically, availability was not a limiting factor, and SCCP returned a complete threshold schedule across all main and sensitivity experiments.

B Additional Details on the Coverage Relation

B.1 Geometry of the Coverage Relation

Figure A6 visualizes the estimated coverage relation $\mathcal{C}_t(u, v)$. Recall that u controls the treatment policy, whereas v is the cutoff used to evaluate coverage. Holding u fixed, increasing v enlarges the prediction set and therefore increases coverage. Along the diagonal $u = v$, however, the same change also modifies the treatment policy and the outcome distribution. Coverage along this diagonal is therefore not generally monotone, unlike in ordinary fixed-policy conformal calibration.

To gain a closer view of this behavior, we visualize the feasible set on MIMIC-CXR, as shown in Figure A7. The shaded regions indicate values of q for which the estimated self-consistent coverage falls below the target. Unlike ordinary fixed-policy calibration, these regions need not form a single interval, since changing q also changes the induced treatment process. This illustrates why the self-consistent coverage constraint cannot generally be solved by ordinary quantile inversion.

B.2 Validation of Coverage Identification

Since SCCP recovers deployment coverage from historical trajectories through prefix weighting, we examine both the accuracy of this identification and the effective information available after weighting. Figure A8 (a) reports the RMSE between coverage estimated from historical trajectories and coverage under the corresponding controlled deployment process. Full prefix weighting yields low estimation error across all four clinical benchmarks, whereas unweighted historical coverage can differ substantially from deployment coverage. Partial weighting schemes perform well on some individual datasets, but their accuracy varies across benchmarks. The full prefix estimator, by contrast, remains consistently accurate across the four settings.

Figure A8 (b) reports the effective sample fraction ESS/N. This quantity summarizes the concentration of the importance weights: values closer to one indicate that a larger fraction of the calibration sample contributes effectively to the estimate, whereas smaller values indicate stronger weight concentration.

The effective sample fraction generally decreases with stage as additional treatment-probability ratios accumulate in the prefix weight. The decline is more pronounced on MIMIC-IV and MIMIC-CXR, while eICU and INSPIRE retain larger effective sample fractions, indicating less concentrated weights and more stable coverage estimation at later stages.

The rate of this decline depends on both length of the treatment prefix and the overlap between the historical and threshold-induced treatment processes. This provides a finite-sample explanation for the increased calibration-data requirements at later stages: as the prefix weights become more concentrated, more historical trajectories are needed to estimate deployment coverage with comparable accuracy.

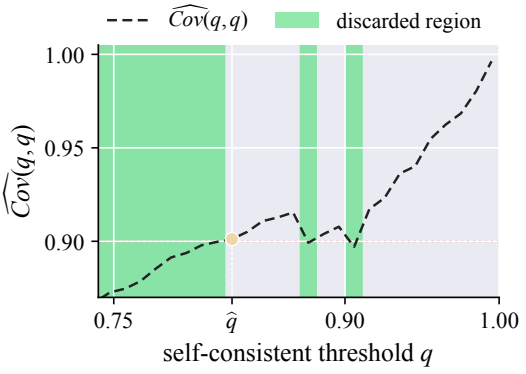

Figure A7: **Diagonal coverage and feasible thresholds on MIMIC-CXR.** The dashed curve shows estimated self-consistent coverage along $u = v = q$, and the horizontal line marks the 90% target; shaded regions are infeasible. **Takeaway.** Self-consistent coverage can be nonmonotone in q, yielding feasible regions that are not characterized by a single quantile threshold.

C Algorithm and Computational Complexity

Algorithm 1 summarizes the calibration procedure of SCCP. The prediction model, historical policy estimate, treatment-policy family, and threshold sets are fixed before the final calibration step. At deployment, the calibrated schedule is fixed. Given the current state–action pair (S_t, A_t) at stage t, the predictive output is $C_{\widehat{q}_t}(S_t, A_t)$.

Computational complexity. Let $K_t = |\mathcal{G}_t|$ denote the number of threshold values evaluated at stage t. For a fixed threshold, computing the importance-weighted coverage and prediction-set size requires a pass over the N calibration trajectories. The previously selected prefix weights can be carried forward from one stage to the next, so evaluating all thresholds at stage t costs $O(NK_t)$, up to the cost of evaluating the treatment policy. The total calibration cost is therefore $O\left(N \sum_{t=0}^{T-1} K_t\right)$. When $K_t = K$ for all stages, this reduces to $O(NTK)$.

The procedure follows a single selected threshold prefix rather than jointly enumerating all possible schedules. A direct enumeration over the complete threshold space would involve $\prod_{t=0}^{T-1} K_t$ possible schedules. The stagewise construction therefore scales linearly with the treatment horizon for fixed K_t, rather than exponentially in the number of stages.

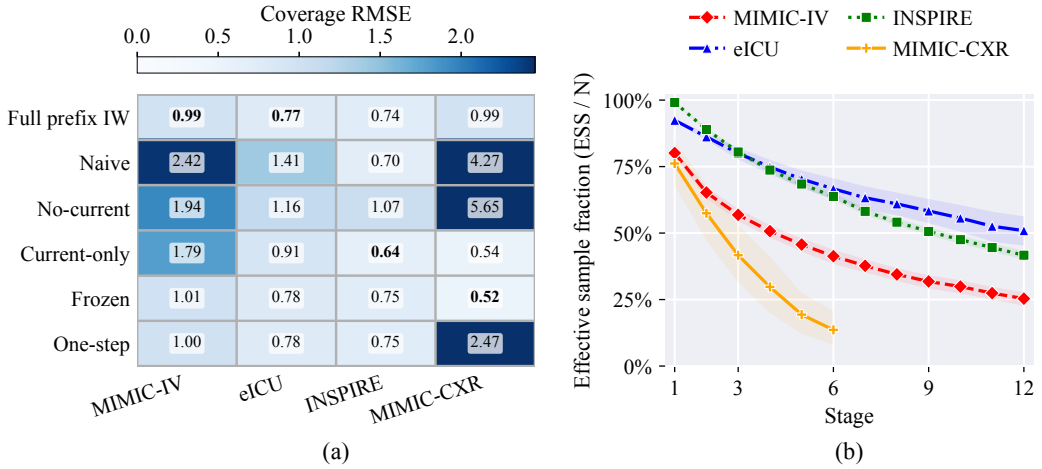

Figure A8: **Validation of coverage identification.** **(a)** Coverage-estimation RMSE, in percentage points, across the four clinical benchmarks. **(b)** Effective sample fraction ESS/N across treatment stages under full prefix weighting. **Takeaway.** Full prefix weighting gives consistently low coverage-identification error, while the effective information available to the estimator decreases as prefix weights accumulate.

Algorithm 1 Self-Consistent Calibration for SCCP

Require: Calibration trajectories $\mathcal{D}_{\mathrm{cal}} = \{\tau_i\}_{i=1}^{N}$, scores $\{R_{i,t}\}$, historical policy $\widehat{\mu}$, treatment-policy family $\{\pi_{t,u}\}$, threshold sets $\{\mathcal{G}_t\}_{t=0}^{T-1}$, target miscoverage level α
Ensure: Calibrated threshold schedule $\widehat{\boldsymbol{q}} = (\widehat{q}_0, \ldots, \widehat{q}_{T-1})$

```
Initialize q̂_{0:-1} = ∅
for t = 0, ..., T − 1 do
    for all u ∈ G_t do
        Compute prefix weights {Ŵ_{i,t}(u; q̂_{0:t−1})}_{i=1}^N
        Compute self-consistent coverage Ĉ_{N,t}(u, u; q̂_{0:t−1})
        Compute weighted prediction-set size Ŝize_{N,t}(u; q̂_{0:t−1})
    end for
    Define the feasible set
        F̂_t ← { u ∈ G_t : Ĉ_{N,t}(u, u; q̂_{0:t−1}) ≥ 1 − α }
    Select  q̂_t ∈ arg min_{u ∈ F̂_t} Ŝize_{N,t}(u; q̂_{0:t−1})
    Append q̂_t to the selected threshold prefix
end for
return q̂ = (q̂_0, ..., q̂_{T−1})
```

Apart from storage of the calibration data and fitted models, maintaining the cumulative prefix weights requires $O(N)$ working memory.

D Implementation and Experimental Details

D.1 Datasets used in this study

We consider one synthetic benchmark and four clinical benchmarks constructed from publicly available healthcare datasets.

Synthetic. We construct a synthetic benchmark with $T = 12$ stages to study conformal calibration under controlled treatment feedback. The benchmark represents a stylized acute-care setting in which an unstable patient's physiological state evolves under repeated treatment decisions. At each stage, the treatment policy chooses among three ordered intervention levels according to the current patient state. Treatment affects subsequent patient difficulty and outcome variability, allowing the direction and strength of prediction-induced feedback to be varied explicitly.

At stage t, the state is $S_t = (Z_{t,1}, Z_{t,2}, D_t, t/T)$, where $Z_{t,1}$ and $Z_{t,2}$ are continuous state variables and $D_t \in \{0, 1\}$ is a binary difficulty indicator. At baseline,

$$(Z_{0,1}, Z_{0,2}) \sim \mathcal{N}(0, I_2), \qquad D_0 \sim \text{Bernoulli}(0.25).$$

There are three treatments, indexed by $a \in \{0, 1, 2\}$, with ordered treatment coordinates $c(a) \in \{-1, 0, 1\}$. The historical policy $\mu(a \mid S_t)$ is state dependent and assigns positive probability to every treatment. We generate it from a softmax model with mild uniform exploration,

$$\mu(a \mid S_t) = 0.88 \frac{\exp\{L_a(S_t)\}}{\sum_{b=0}^{2} \exp\{L_b(S_t)\}} + \frac{0.12}{3},$$

where

$$\begin{aligned} L_0(S_t) &= 0.15 - 0.10Z_{t,1} + 0.15Z_{t,2} - 0.20D_t, \\ L_1(S_t) &= 0.10Z_{t,1} - 0.05Z_{t,2}, \\ L_2(S_t) &= -0.10 + 0.25Z_{t,1} - 0.10Z_{t,2} + 0.35D_t. \end{aligned}$$

To introduce performative effects, we let the deployment policy vary with the conformal threshold q. We first map q smoothly to $[0, 1]$,

$$r(q) = \frac{\text{expit}\big(4[\bar{q} - \frac{1}{2}]\big) - \text{expit}(-2)}{\text{expit}(2) - \text{expit}(-2)}, \qquad \bar{q} = \left[\frac{q - 1.00}{1.80 - 1.00}\right]_{[0,1]}.$$

where $\text{expit}(x) = (1 + e^{-x})^{-1}$ denotes the logistic sigmoid function.

The deployment policy then tilts the historical policy toward treatments with smaller $c(a)$ as q increases,

$$\pi_q(a \mid S_t) = \mu(a \mid S_t) \min\left\{\frac{\exp\{-0.4r(q)c(a)\}}{z(S_t, q)}, 3\right\},$$

where $z(S_t, q)$ is chosen so that the resulting probabilities sum to one. Hence $q \leq 1.00$ recovers the historical policy, while larger thresholds induce progressively greater changes in treatment. The factor of three bounds the policy shift itself. The cumulative importance weights used by SCCP are left unclipped.

Treatment affects future patient difficulty through

$$H_t = \gamma c(A_t) D_t,$$

where γ controls the strength and direction of feedback. The next-stage difficulty indicator is generated by

$$D_{t+1} \sim \text{Bernoulli}\left[\text{expit}\left\{-1.60 + 2.00D_t + 0.25\tanh(Z_{t,1}) + H_t\right\}\right].$$

We also introduce a rare high-variance event,

$$B_{t+1} \sim \text{Bernoulli}\left[\text{expit}\left\{-3.20 + 2.20D_t + 1.50H_t\right\}\right].$$

The continuous state evolves as

$$\begin{aligned} Z_{t+1,1} &= 0.78Z_{t,1} + 0.12Z_{t,2} + 0.30D_{t+1} + 0.25\eta_{t,1}, \\ Z_{t+1,2} &= 0.08Z_{t,1} + 0.72Z_{t,2} + 0.20D_{t+1} + 0.20\eta_{t,2}, \end{aligned}$$

where $(\eta_{t,1}, \eta_{t,2})$ is bivariate Gaussian noise with correlation 0.30. The two-dimensional outcome is

$$Y_{t+1} = \begin{pmatrix} Z_{t+1,1} \\ Z_{t+1,2} \end{pmatrix} + \begin{pmatrix} 0.25 \\ 0.20 \end{pmatrix} D_{t+1} + \{1 + 5B_{t+1}\} \begin{pmatrix} 0.35\varepsilon_{t,1} \\ 0.25\varepsilon_{t,2} \end{pmatrix},$$

where $(\varepsilon_{t,1}, \varepsilon_{t,2})$ is independent bivariate Gaussian noise with correlation 0.30. Hence, difficult states increase both the mean response and the probability of unusually large prediction errors.

In this construction, negative γ makes the threshold-induced policy shift increase subsequent difficulty and the probability of large prediction errors, while positive γ reverses this direction. Setting $\gamma = 0$ removes this treatment feedback interaction. Our primary experiments use $\gamma = -4$.

Table A2: Sequential treatment tasks constructed from the four clinical datasets. Responses are two-dimensional burdens measured after the corresponding action window.

Dataset	Episodes	Horizon	Response pair	Action space
MIMIC-IV	22,416	12×4 h	Hypotension and tachycardia burden	Combined fluid and vasopressor treatment levels
eICU	6,175	12×4 h	Hypotension and tachycardia burden	Combined fluid and vasopressor treatment levels
INSPIRE	45,117	12×10 min	Hypotension and hypertension burden	No treatment, fluid only, or vasopressor treatment
MIMIC-CXR	1,993	6×6 h	Hypoxemia and tachypnea burden	No respiratory support, non-invasive support, or invasive ventilation

MIMIC-IV. MIMIC-IV is a large de-identified electronic health record database collected at Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al., 2024; 2023). It contains hospital and intensive-care data from patients treated between 2008 and 2022, including demographics, admissions, laboratory measurements, medications, procedures, fluid inputs, outputs, and continuously charted clinical variables. Its ICU module provides detailed time-stamped measurements and treatment records, making it well suited to constructing longitudinal treatment trajectories.

eICU. The eICU Collaborative Research Database is a de-identified, multi-center critical-care database collected through the Philips eICU program (Pollard et al., 2018). It contains more than 200,000 ICU encounters from over 139,000 patients admitted to 335 units at 208 hospitals across the United States during 2014–2015. Available data include vital signs, laboratory measurements, medications, treatment records, diagnoses, care plans, and APACHE severity measures. Compared with the single-center MIMIC-IV database, eICU provides a geographically and institutionally broader critical-care population.

INSPIRE. INSPIRE is a de-identified perioperative dataset from Seoul National University Hospital in South Korea (Lim et al., 2024; Lim & Lee, 2026). It contains approximately 130,000 surgical cases from patients who underwent anesthesia between 2011 and 2020. The database includes patient characteristics, diagnoses, surgical procedures, anesthesia information, intraoperative and ward vital signs, laboratory measurements, medications, and postoperative outcomes. Intraoperative vital signs and anesthesia-machine variables are recorded at high temporal resolution, making INSPIRE particularly suitable for studying short-horizon treatment decisions during surgery.

MIMIC-CXR. MIMIC-CXR is a large chest radiograph database collected at Beth Israel Deaconess Medical Center (Johnson et al., 2019a;b). It contains 377,110 chest radiographs from 227,835 imaging studies, together with free-text radiology reports. The images are provided in DICOM format and preserve the within-patient temporal ordering required for longitudinal analysis. MIMIC-CXR can also be linked to the corresponding MIMIC clinical records through patient identifiers.

Clinical task construction. We now describe the processing of each dataset in detail. Table A2 summarizes the four controlled clinical benchmarks. Each task is formulated as a sequential treatment problem with two post-treatment outcomes at each stage.

For each clinical dataset, we construct a longitudinal treatment task using only information available before treatment. Each interval comprises an action window followed by a response window, ensuring that treatment precedes the outcomes used for evaluation.

For MIMIC-IV and eICU, eligible ICU stays are divided into 12 four-hour stages. Treatment is defined by joint fluid and vasopressor use, with subsequent hypotension and tachycardia burden as the two outcomes. Fluid and vasopressor exposures are discretized using medians estimated from the prediction-model split, and rare treatment combinations are merged with the nearest sufficiently supported category. INSPIRE comprises 12 ten-minute stages, with treatment defined by fluid and

Table A3: Comparison of the information requirements and calibration structure of the evaluated methods. "Multistep feedback" indicates whether the method explicitly accounts for feedback accumulated over multiple rounds, while "treatment-induced state shift" indicates whether earlier treatment decisions are allowed to alter the state distribution at subsequent treatment stages.

Method	Post-deployment outcomes?	Calibration rule	Score / set	Multistep feedback?	Treatment-induced state shift?
Standard CP	No	Independent split-conformal quantile at each stage	normalized; box	No	No
ACI	Yes	Online update of the nominal miscoverage level from sequential coverage errors	normalized; dynamic box	No	No
MFCS	No	Adapted finite-history feedback-shift weighting	normalized; box	Yes	No
PRC	Yes	Sequential performative-risk calibration using repeated target-policy cohorts	normalized; box	-	Yes
SCCP	**No**	Stagewise self-consistent calibration using full prefix weights	normalized; box	**Yes**	**Yes**

vasopressor use and subsequent hypotension and hypertension burden as outcomes. MIMIC-CXR comprises six six-hour stages, with respiratory support level as treatment and subsequent hypoxemia and tachypnea burden as outcomes. Treatment categories for INSPIRE and MIMIC-CXR are specified before patient splitting.

D.2 Additional experimental setup

Baselines and implementations. Our evaluation asks whether stagewise coverage under a threshold-induced treatment process can be calibrated from historical trajectories before deployment. We use standard split conformal prediction (Standard CP) as the primary feedback-agnostic baseline: it receives the same historical calibration trajectories as SCCP but does not adjust for the treatment process induced by its selected thresholds.

We complement this comparison with references from related calibration approaches: Adaptive Conformal Inference (ACI) (Gibbs & Candes, 2021), Multi-step Feedback Covariate Shift (MFCS) (Prinster et al., 2024), and Performative Risk Control (PRC) (Li et al., 2025). They provide reference points for online adaptation, multistep feedback-shift correction, and deployment-based risk calibration, respectively. Table A3 summarizes their roles and calibration procedures.

For all methods producing rectangular prediction sets, we use the same fitted outcome model and the same scale-normalized nonconformity score,

$$R_t = \max_{j\in\{1,2\}} \frac{|Y_{t+1,j} - \widehat{m}_j(S_t, A_t)|}{\widehat{\sigma}_j(S_t, A_t)}.$$

For threshold q, the corresponding rectangular prediction set is

$$C_q(s,a) = \prod_{j=1}^{2} \left[\widehat{m}_j(s,a) - q\widehat{\sigma}_j(s,a), \widehat{m}_j(s,a) + q\widehat{\sigma}_j(s,a)\right].$$

Table A4: Baseline implementations used in our experiments.

Method	Implementation
ACI	https://github.com/isgibbs/AdaptiveConformal
MFCS	https://github.com/drewprinster/conformal-mfcs
PRC	https://github.com/livctr/rcpp

Thus, differences among Standard CP, ACI, MFCS, PRC, and SCCP arise from their calibration procedures and information available during calibration, rather than from different predictive models or prediction-set geometries.

Standard CP. At each treatment stage, Standard CP applies ordinary split-conformal calibration to the historical calibration scores and selects the corresponding conformal quantile independently across stages. It therefore calibrates under the historical treatment process and does not account for threshold-induced treatment feedback.

ACI. ACI retains its online calibration update. Let α_k denote its current miscoverage level and err_k the coverage error observed after deployment. We use

$$\alpha_{k+1} = \alpha_k + \gamma_{\mathrm{ACI}} \left(\alpha - \mathrm{err}_k\right), \qquad \gamma_{\mathrm{ACI}} = 0.01.$$

ACI receives target-policy outcomes in chronological order and updates the corresponding rectangular prediction sets as these outcomes become available.

MFCS. MFCS extends conformal prediction to multistep feedback covariate shift, where the data-collection process can affect the distribution of future observations. To apply it to our sequential treatment setting, we represent this feedback through treatment-assignment probabilities. Our implementation uses a history depth of three, caps the cumulative weight at 40, and calibrates a single scale multiplying a fixed stage profile. These choices are fixed across all datasets and settings.

Native PRC. PRC (Li et al., 2025) controls performative risk through repeated observations from the deployment process. We first retain its native trajectory-level formulation as a reference. The loss records whether a trajectory contains any miscovered stage, and PRC iteratively updates its calibration parameter using repeated target-policy cohorts. We use $(\alpha, \delta) = (0.10, 0.05)$, tightness 0.08, $\tau = 1$, and target-policy cohorts of size $2{,}000$.

Stage-wise PRC. To align PRC with the stagewise marginal coverage target used by SCCP, we apply its performative-risk update sequentially across treatment stages. At stage t, let $\widehat{\boldsymbol{q}}^{\mathrm{PRC}}_{0:t-1}$ denote the thresholds already selected by PRC. Holding this prefix fixed, define

$$\mathcal{D}^{\mathrm{PRC}}_t(u) = P_t\left(\widehat{\boldsymbol{q}}^{\mathrm{PRC}}_{0:t-1}, u\right),$$
$$\mathcal{R}^{\mathrm{PRC}}_t(u, v) = \mathbb{E}_{Z_t \sim \mathcal{D}^{\mathrm{PRC}}_t(u)} \left[\mathbf{1}\{r(Z_t) > v\}\right],$$

where $Z_t = (S_t, A_t, Y_{t+1})$. The calibration target is

$$\mathcal{R}^{\mathrm{PRC}}_t \left(\widehat{q}^{\mathrm{PRC}}_t, \widehat{q}^{\mathrm{PRC}}_t\right) \le \alpha, \qquad t = 0, \ldots, T-1.$$

Thus, Stage-wise PRC targets the same stagewise marginal miscoverage level as SCCP, rather than the probability of any miscoverage event over an entire trajectory. Because controlling any-stage miscoverage is a stronger requirement, we retain Native PRC only as a supplementary reference. The main PRC comparison uses the stagewise formulation above.

In the stagewise implementation, we use $\alpha = 0.10$, tightness 0.08, $\tau_{\mathrm{PRC}} = 1$, and target-policy cohorts of size $2{,}000$ per deployment iteration. The confidence parameter $\delta_{\mathrm{PRC}} = 0.05$ is applied separately at each stage; accordingly, we do not interpret it as a simultaneous 95% confidence guarantee over the full treatment horizon. Thus, Stage-wise PRC provides a deployment-based comparison for the same stagewise marginal coverage target, whereas SCCP selects its threshold schedule using historical-policy trajectories only.

The public repositories in Table A4 serve as the reference implementations for the corresponding methods. Our implementation is available in the anonymous repository at https://anonymous.4open.science/r/sccp, and additional supplementary materials are provided with the anonymous submission.

Data splits and evaluation. Patients are split into three non-overlapping groups. The first is used to fit the outcome model and the historical treatment policy. The second is used to define how the conformal threshold changes treatment. The third is used to fit the controlled transition environment for semi-synthetic evaluation. For MIMIC-IV, eICU, and INSPIRE, the three splits contain 40%, 20%, and 40% of the patients, respectively. For MIMIC-CXR, the proportions are 20%, 20%, and 60%. All fitted components are fixed before calibration. Afterward, all pilot, calibration, and evaluation trajectories are generated from the fixed controlled environment.

For each dataset, seed, and feedback setting, an independent pilot sample is first generated under the fitted historical policy and used only to construct the threshold grid at each stage. Each grid contains 101 empirical score quantiles between 0.50 and 0.999 and is fixed before calibration. The quantile grid adapts to the stage-specific score scale and allows the potentially nonmonotone coverage constraint to be evaluated directly. We then generate an independent calibration sample under the same historical policy.

Standard CP, MFCS, and SCCP use the same calibration trajectories, while ACI uses the corresponding historical-policy data for initialization before its online updates. PRC instead uses target-policy cohorts generated during its stagewise calibration procedure. The pilot trajectories are not reused for calibration. After calibration or online adaptation, each method is evaluated on fresh target-policy trajectories that are not used for threshold selection or adaptation.

Prediction models. The clinical outcome and historical treatment models are two-layer GRUs using the four most recent states, with hidden dimension 128 and representation dimension 32. The outcome models return a two-dimensional conditional mean together with positive coordinate-wise scales. For MIMIC-CXR, a DenseNet-121 initialized from ImageNet is fine-tuned using only patients in the prediction-model split. Its image representation is then frozen and combined with the longitudinal clinical features. The Synthetic uses a multilayer perceptron in place of the clinical sequence model. All prediction models are fixed before conformal calibration.

Metrics. For seed s, let $\widehat{C}_{s,t}$ denote empirical coverage at stage t. For methods with rectangular prediction sets,

$$\widehat{C}_{s,t} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{1}\left\{Y_{s,i,t+1} \in C_{q_{s,i,t}}(S_{s,i,t}, A_{s,i,t})\right\},$$

where $q_{s,i,t}$ denotes the threshold applied to patient i at stage t. Let $\mathcal{R}$ denote the prespecified set of random seeds, and let $\mathcal{S} \subseteq \mathcal{R}$ denote the runs in which the method produces its required output. For SCCP, this requires a complete threshold schedule.

For each run, we define

$$\mathrm{WSC}_s = \min_{0 \le t < T} \widehat{C}_{s,t}, \qquad \mathrm{MeanCov}_s = \frac{1}{T} \sum_{t=0}^{T-1} \widehat{C}_{s,t}.$$

We report WSC and MeanCov by averaging these quantities across runs.

We measure efficiency by the run-level mean normalized coordinate width,

$$\mathrm{NWidth}_s = \frac{1}{N_s T} \sum_{i=1}^{N_s} \sum_{t=0}^{T-1} \frac{1}{2} \sum_{j=1}^{2} \frac{2 q_{s,i,t} \widehat{\sigma}_{s,i,t,j}}{\sigma_{s,j}^{\mathrm{out}}},$$

where $\sigma_{s,j}^{\mathrm{out}}$ is the sample standard deviation of outcome j in the prediction-model subset. Smaller NWidth indicates more informative prediction sets.

Treatment policy and decision feedback. For each clinical benchmark, let $\mu_t(a \mid s)$ denote the fitted historical treatment policy and let $\pi_t^{\mathrm{alt}}(a \mid s)$ denote a fixed alternative policy. We construct π_t^{alt} from the policy-design data by reweighting the historical policy toward treatments with lower predicted clinical and treatment costs. To limit the policy shift and maintain overlap with the historical data, we constrain

$$\frac{\pi_t^{\mathrm{alt}}(a \mid s)}{\mu_t(a \mid s)} \le 3$$

for all supported state–action pairs. The resulting policy is then used directly when computing the prefix importance weights.

We let q_L and q_U be the 0.80 and 0.95 quantiles of the pooled nonconformity scores in the independent policy-design subset. For a threshold q, define

$$\widetilde{q} = \mathrm{clip}\left(\frac{q - q_L}{q_U - q_L}, 0, 1\right), \qquad \lambda(q) = \frac{\ell\{4(\widetilde{q} - 1/2)\} - \ell(-2)}{\ell(2) - \ell(-2)}, \qquad \ell(z) = \frac{1}{1 + e^{-z}}.$$

The treatment policy associated with q is

$$\pi_{t,q}(a \mid s) = \{1 - \lambda(q)\}\mu_t(a \mid s) + \lambda(q)\pi_t^{\mathrm{alt}}(a \mid s).$$

Thus, $q \le q_L$ recovers the historical treatment policy, whereas $q \ge q_U$ reaches the fixed alternative policy, with intermediate thresholds smoothly interpolating between the two.

For the clinical benchmarks, these policies are evaluated in a controlled semi-synthetic environment constructed only from the held-out environment subset $\mathcal{D}_{\mathrm{env}}$. At each stage, a ridge model predicts the next clinical state from the current state representation and treatment. The remaining transition variation is supplied by a residual drawn from a stage- and treatment-specific donor library. Donors are selected from the 100 nearest state representations using a Gaussian kernel with bandwidth 2. If donor J is selected, its state residual is added to the ridge prediction, while the outcome is generated as

$$Y_{t+1} = \widehat{m}(S_t, A_t) + \widehat{\sigma}(S_t, A_t)\widehat{\varepsilon}_{J,t},$$

where $\widehat{\varepsilon}_{J,t}$ is the standardized outcome residual of donor J under the frozen outcome model.

Treatment feedback is introduced through the donor-selection probabilities. For a candidate donor j under action a, the sampling weight is proportional to

$$\exp\left\{-\frac{\widetilde{d}_j^2}{2h^2} + \gamma\, c(a) D_{j,t}\right\},$$

where $\widetilde{d}_j$ is the locally normalized state distance, $h = 2$ is the kernel bandwidth, $c(a)$ is the prespecified action coordinate, and $D_{j,t}$ is the donor difficulty score. Hence, γ controls the strength and direction of the treatment-induced shift in subsequent states and outcomes. We use $\gamma = -4$ as the primary setting and $\gamma \in \{-2, 0, 2, 4\}$ for sensitivity analysis.

For a fixed dataset and γ, all methods share the same initial-state distribution and transition–outcome kernel $K_t^{(\gamma)}$; they differ only through the treatment policy used to generate actions. In particular, the conformal threshold does not enter the transition kernel directly, so its effect on future states and outcomes is mediated entirely through treatment. Pilot and final calibration trajectories are generated under the fitted historical policy μ. Full implementations of the treatment-policy family and controlled clinical environment are provided with the code release.

E Additional Experimental Results

This section provides additional experiments that complement the main results. We first visualize the self-consistent coverage relation, then examine the effects of calibration sample size and treatment horizon, and finally provide additional baseline comparisons and ablation studies.

E.1 Sensitivity to Calibration Size and Treatment Horizon

We next examine how the amount of calibration data interacts with the length of the treatment process. Overall, longer treatment horizons require more calibration data to maintain stable worst-stage coverage, but this sensitivity diminishes as the calibration sample grows.

Figures A9 and A10 show a consistent pattern across the four clinical benchmarks. Specifically, with limited calibration data, increasing the treatment horizon can make worst-stage coverage less stable. As the calibration sample grows, worst-stage coverage moves toward the nominal level and the differences between treatment horizons become smaller. The same qualitative pattern appears across the four clinical benchmarks. This behavior is consistent with the sequential weighting used by

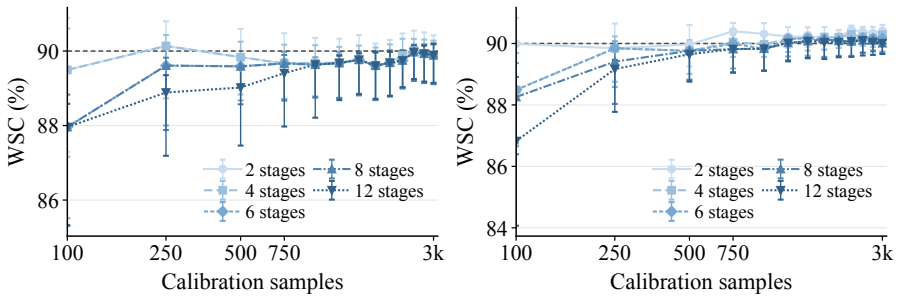

Figure A9: **Calibration-size sensitivity under increasing treatment horizon for MIMIC-IV and eICU.** Worst-stage coverage is shown across calibration sample sizes for different treatment horizons; the dashed line marks the 90% target. **Takeaway.** Longer horizons require more calibration data, but the horizon gap shrinks as the calibration sample grows.

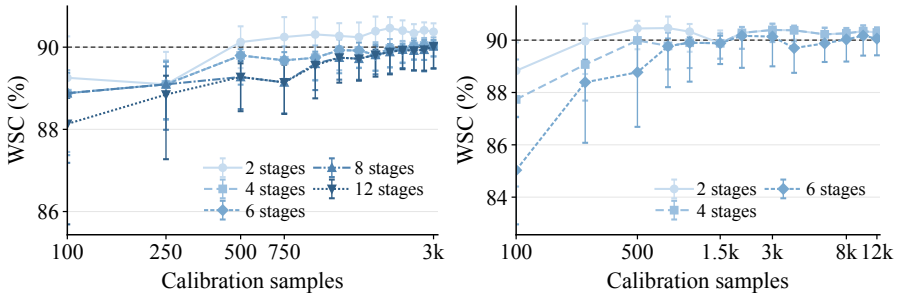

Figure A10: **Calibration-size sensitivity under increasing treatment horizon for INSPIRE and MIMIC-CXR. Takeaway.** Coverage becomes increasingly insensitive to treatment horizon as more calibration trajectories are available.

SCCP. At later stages, the coverage estimator depends on a longer sequence of treatment-probability ratios. Estimating the resulting target distribution is therefore more difficult when only a small number of historical trajectories are available. The experiments show, however, that the effect of treatment horizon becomes much smaller as the calibration sample increases.

Stagewise Stability with Increasing Calibration Size. Worst-stage coverage summarizes performance by the most difficult treatment stage. We additionally examine the full stagewise coverage profile. Figures A11 and A12 provide a stage-by-stage view of the same effect. With small calibration samples, coverage can fluctuate noticeably across treatment stages. These fluctuations become smaller as the number of calibration trajectories increases, and the stagewise profiles concentrate more closely around the target level.

The improvement is not confined to a single favorable stage. Rather, additional calibration data improve stability across the treatment horizon as a whole. We also do not observe a systematic deterioration at later stages once the calibration sample is sufficiently large in these experiments.

Baseline Sensitivity to Calibration Size. We also vary the calibration sample size for all compared methods to examine whether the main comparison depends on a particular choice of calibration-set size.

Figure A13 shows that the qualitative comparison in the main experiments is not specific to one calibration sample size. Standard CP remains below the nominal target under treatment feedback, while ACI improves coverage but can still fall below the target. MFCS tends to remain above the target, whereas SCCP stays close to the nominal level as the calibration sample changes. PRC shows substantially different behavior across the benchmarks, reflecting its different deployment-based adaptation dynamics and information regime.

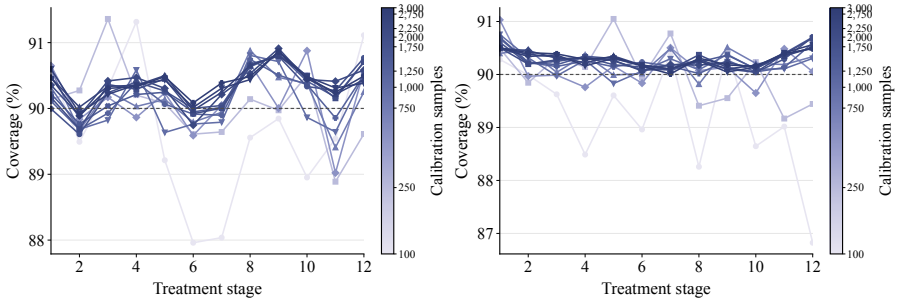

Figure A11: **Stagewise coverage versus calibration sample sizes for MIMIC-IV and eICU.** Each curve shows coverage across treatment stages for a different calibration sample size; darker curves correspond to larger samples. **Takeaway.** More calibration data reduce stage-to-stage variation and stabilize coverage around the 90% target.

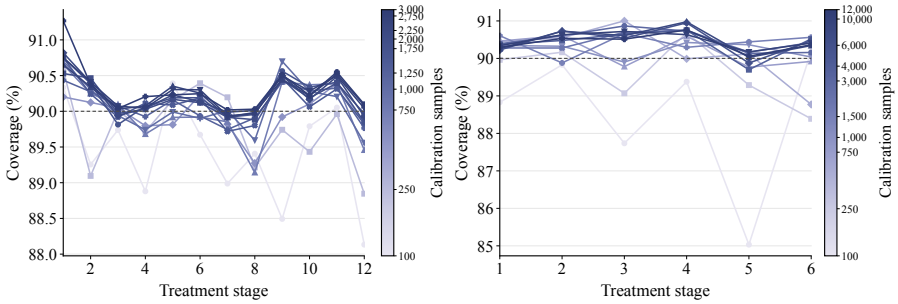

Figure A12: **Stagewise coverage versus calibration sample sizes for INSPIRE and MIMIC-CXR. Takeaway.** Larger calibration samples produce more stable coverage across the full treatment horizon.

Figure A14 shows the corresponding prediction-set widths. The relative efficiency patterns are also broadly stable as more calibration data are added. In particular, the conservative coverage of MFCS is generally accompanied by wider prediction sets, while the width of SCCP remains comparatively stable. Together, these results indicate that the main comparison is not an artifact of a particular calibration sample size.

E.2 Additional Ablation Results

We extend the ablation study in the main text to the remaining benchmarks. All variants use the same data, prediction model, score, candidate thresholds, and evaluation protocol as SCCP. In particular, Frozen removes the dependence of the treatment process on the current candidate threshold, One-step omits recursive treatment-order calibration, and No-current and Current-only remove complementary parts of the prefix likelihood ratio.

Across the benchmarks, the ablations show that both parts of the prefix correction can matter. Removing the current treatment ratio ignores the change in treatment received before observing Y_{t+1}, whereas using only the current ratio ignores the earlier treatment decisions that shape the state distribution encountered at stage t. The effects are most visible under stronger feedback, where the induced treatment process differs more substantially from the historical process.

The width results provide an important qualification. Some ablated variants produce narrower prediction sets in particular settings, but these reductions do not consistently preserve worst-stage coverage. Thus, the full method does not simply enlarge prediction sets to recover coverage; rather, the results support accounting for both the current treatment and earlier treatment decisions when calibrating under sequential feedback.

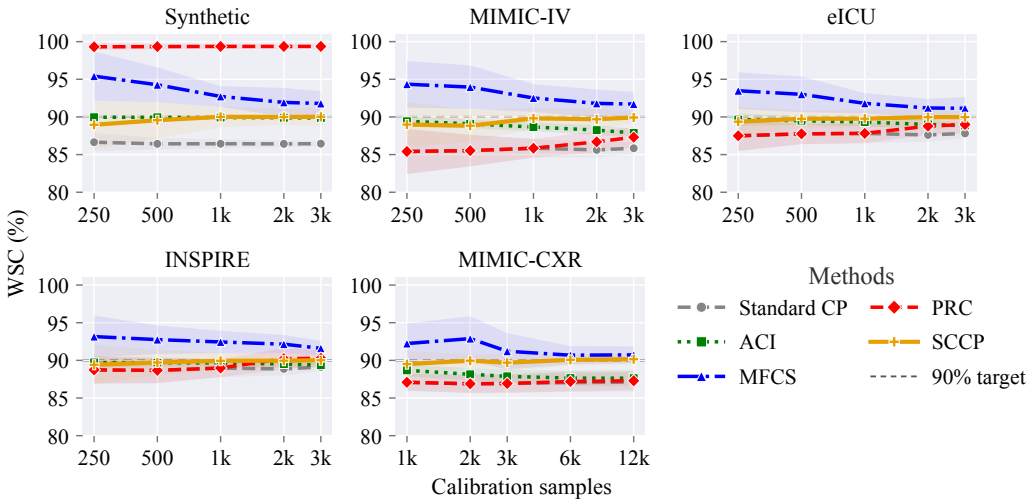

Figure A13: **Worst-stage coverage versus calibration sample sizes.** Worst-stage coverage is reported for Standard CP, ACI, MFCS, PRC, and SCCP across the five benchmarks. **Takeaway.** SCCP remains close to the nominal target across calibration sizes, while the relative coverage behavior of the baselines remains broadly unchanged.

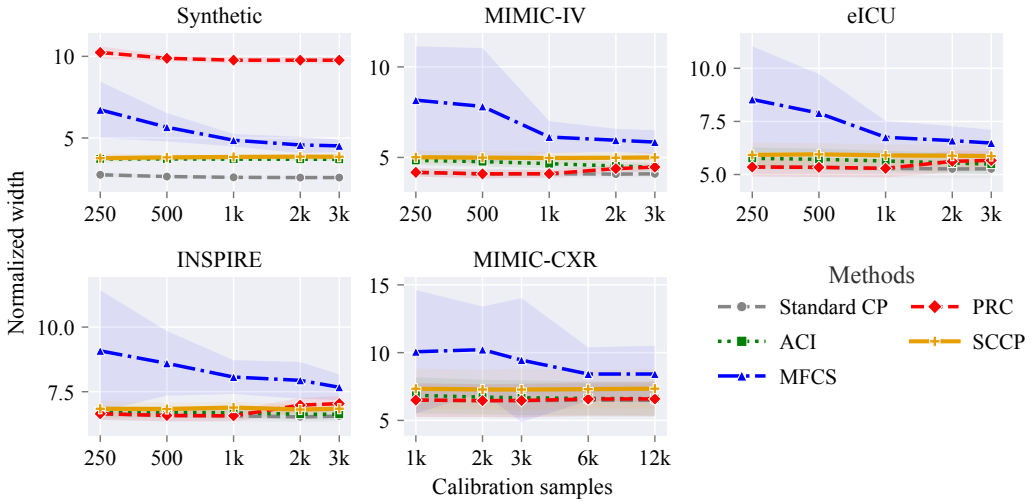

Figure A14: **Prediction-set width versus calibration sample size.** Normalized prediction-set width is shown for the methods across the five benchmarks. **Takeaway.** The relative width patterns are stable across calibration sizes, with MFCS generally producing wider sets than SCCP.

F Supplement to Related Work

Performative prediction. The broader performative prediction literature studies learning problems in which deploying a predictor changes the distribution on which it is subsequently evaluated. Perdomo et al. (2020) formalize this dependence through a decision-dependent distribution map and introduce performative stability as an equilibrium notion. Subsequent work studies how such stable solutions can be reached through stochastic optimization, including the trade-off between updating and redeploying the model (Mendler-Dünner et al., 2020). Subsequent work extends this view beyond a one-step response. In particular, Brown et al. (2022) consider stateful environments in

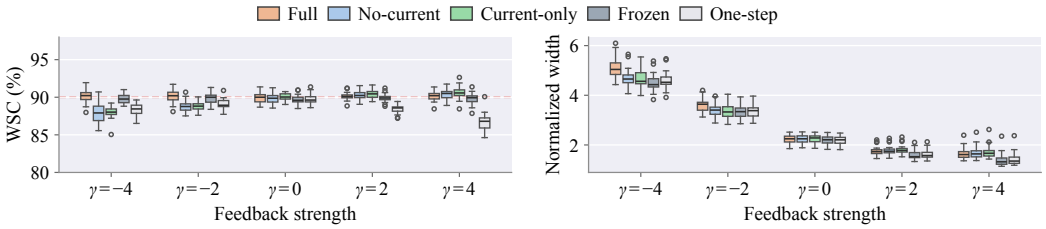

Figure A15: **Ablation results on the Synthetic.** Boxplots show worst-stage coverage (left) and normalized prediction-set width (right) across feedback strengths. **Takeaway.** The full method maintains coverage near the target across feedback strengths without relying on uniformly larger prediction sets.

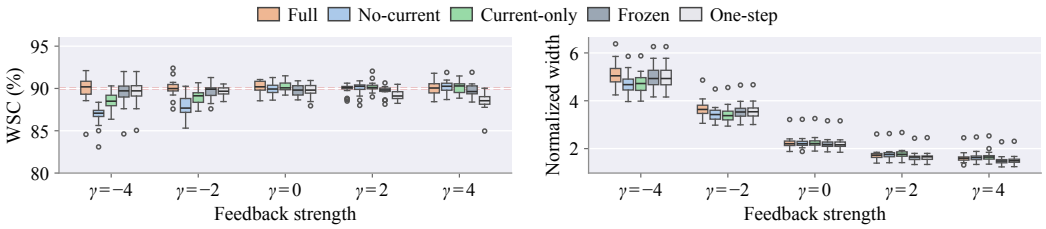

Figure A16: **Ablation results on MIMIC-IV. Takeaway.** Removing the current or preceding treatment correction can produce narrower sets, but the gain is accompanied by lower worst-stage coverage under stronger feedback.

which the population response depends jointly on the deployed predictor and the current population state, so repeated decisions can change the states encountered over time. These formulations are closely related to sequential treatment feedback, since current decisions can shape the distribution encountered at later stages.

This perspective provides a natural starting point for our problem, but the object of interest is different. We keep the predictive model fixed and instead calibrate the conformal threshold, which both defines the prediction set and can influence treatment. Coverage must therefore be evaluated under the treatment process induced by the use of that threshold. Our goal is not to optimize or retrain a predictor toward a stable solution, but to calibrate prediction sets whose coverage remains reliable under the treatment process they help induce. This perspective motivates the self-consistent coverage target studied in our work.

Conformal prediction under distribution shift. A broad line of work extends conformal prediction beyond the exchangeable setting by adapting calibration to changes in the data distribution. For sequential observations, EnbPI uses recent residual information to construct prediction intervals for dynamic time series (Xu & Xie, 2021), while SPCI estimates the conditional quantile of a future nonconformity score from past residuals (Xu & Xie, 2023). MultiDimSPCI further extends this approach to multivariate outcomes using ellipsoidal prediction regions (Xu et al., 2024). More general approaches address non-exchangeability and covariate shift through weighted calibration, transporting information from an observed distribution to a target distribution (Gibbs & Candès, 2024; Tibshirani et al., 2019). Related ideas also appear in conformal inference under feedback covariate shift, where the sampling process may depend on earlier observations (Prinster et al., 2024).

This literature is closely connected to the identification step in SCCP. Historical trajectories follow P_μ, whereas the coverage of interest is defined under a different treatment process $P_t(\boldsymbol{q}_{0:t})$. The prefix likelihood ratio therefore plays the same basic role as importance weighting under covariate or policy shift: it transports coverage from the historical distribution to the treatment process of interest. The additional difficulty is that this target process is not fixed independently of calibration, since

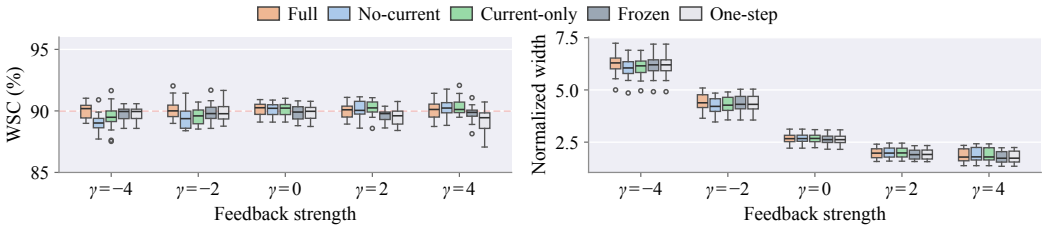

Figure A17: **Ablation results on eICU. Takeaway.** Differences between variants are smaller on eICU, while the full method remains consistently close to the nominal coverage level.

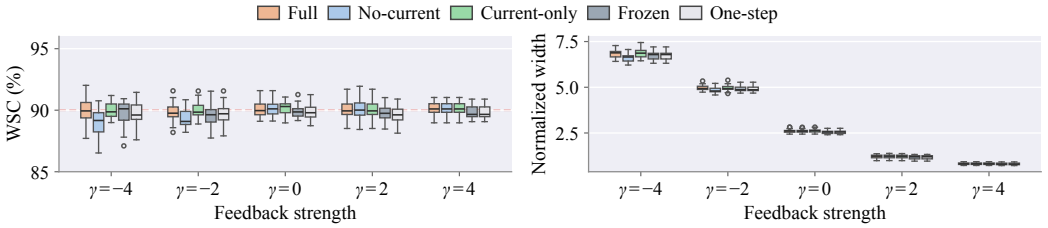

Figure A18: **Ablation results on INSPIRE. Takeaway.** The full construction provides the most stable coverage across feedback strengths, while removing parts of the treatment-history correction offers no consistent coverage–width improvement.

changing the conformal threshold also changes treatment. SCCP therefore combines distribution-shift correction with the self-consistent coverage target motivated above.

G LIMITATIONS

The present study focuses on establishing self-consistent calibration at the level of stagewise population coverage. This is a natural starting point, but clinical treatment decisions are rarely homogeneous across patients. The same prediction set may lead to different treatment responses depending on disease severity, prior treatment history, comorbidities, or clinician preference. A useful next step is therefore to move beyond a common stagewise calibration rule and study subgroup-specific or patient-adaptive calibration. Such an extension would be particularly relevant for personalized medicine, where reliable uncertainty should reflect not only the treatment stage but also clinically meaningful heterogeneity in how predictions influence decisions.

Moreover, our clinical experiments are based on controlled semi-synthetic environments constructed from real patient data. This design gives us a setting in which treatment feedback can be varied systematically and evaluated under the same underlying patient model across methods. At the same time, it cannot reproduce all aspects of real clinical deployment, including variation in clinician behavior, delayed responses, and unobserved factors that may influence treatment decisions. Prospective studies will therefore be necessary to determine how these effects interact with self-consistent calibration when prediction sets are used directly in clinical practice.