CNCRC: PAIRWISE COST-WEIGHTED SCORE DESIGN FOR SPLIT-CONFORMAL PREDICTION SETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Split-conformal prediction gives marginal coverage for any fixed nonconformity score, but it does not determine which alternatives a prediction set retains. We study CNCRC as an orientation-aware score-design family for fixed-label multiclass prediction with a supplied, bounded pairwise ambiguity-cost matrix. Its row and column forms are dual design choices: the row form supports conditional retained-set accounting, whereas the column form weights candidate admission by incoming cost. MAX and SUM provide peak-sensitive and accumulation-sensitive reductions while leaving the calibration template unchanged. We separate generic coverage and bounded-cost omission from the retained-set accounting properties induced by each score. In a validation-only, matched-coverage clinical benchmark, both orientations yield comparable composition metrics, with the column MAX form attaining AmbCost 0.293 and Hazard 10.1%. The poison stress test and the large-label clinical benchmark are interpreted as operating-point evidence about set composition. A poison-class control with the same recall and average set size shows lower poison-label admission on benign examples than random set inflation. The risk-budget sweep and a synthetic cost-matrix perturbation analysis characterize the operating range and sensitivity of the score family. These results characterize oriented score design for the stated benchmark without claiming a universal safety or downstream-utility guarantee.

1 INTRODUCTION

Uncertainty-aware classifiers should communicate not only whether a label is plausible, but also which retained alternatives may be consequential. Split-conformal prediction (CP) provides a finite-sample marginal coverage guarantee for prediction sets under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2021). The guarantee is deliberately agnostic to the model and to the choice of a fixed nonconformity score. Consequently, coverage and set size alone do not describe the asymmetric consequences of the labels that remain in a set.

We call the proposed score layer *Cost-Normalized Conformal Risk Composition* (CNCRC). Here, "composition" refers to the labels retained in a prediction set: CNCRC modifies score composition before ordinary split-CP calibration, and is not a replacement for Conformal Risk Control (CRC) or a direct set-level risk controller.

In this paper, the pairwise matrix is an ambiguity-cost specification: $c(y, z)$ is the consequence assigned to retaining distractor z when y is the reference label. It is not, by itself, an omission loss for failing to retain y, nor does it specify how a downstream policy acts on the set. This distinction is essential because the standard prediction rule retains candidates with score at most the calibrated threshold. A cost-aware score can therefore change set composition without providing a monotone guarantee that a high-consequence reference label is retained.

We study how this matrix enters the *score-design* layer of split CP. For a candidate label y, define:

$$s_{\max}(x, y) = \max_{z \neq y} \hat{p}_z(x) c(y, z), \qquad s_{\text{sum}}(x, y) = \sum_{z \neq y} \hat{p}_z(x) c(y, z). \tag{1}$$

The first emphasizes a single peak confusion and the second aggregates many potential confusions. The matrix orientation is an explicit design input: the displayed row form uses (c(y,z)), while an

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.

incoming or column form replaces it by (c(z,y)). Once a score type and orientation are fixed, calibration and prediction follow the ordinary split-conformal procedure. Thus, CNCRC defines an auditable row–column family of cost-weighted rankings whose empirical behavior can be compared under the same protocol.

This score-design view complements, rather than replaces, cost-aware conformal methods that optimize a supplied downstream utility or calibrate a specified set-level loss (Cortes-Gomez et al., 2024; Lin et al., 2023; Angelopoulos et al., 2022). Such methods and CNCRC both depend on task-specific information. The distinction is where that information enters: CNCRC uses pairwise costs to define a deterministic label-level score before calibration, whereas a utility- or decision-aware method can directly encode how a downstream policy acts on the set. The present paper therefore asks a narrower question: whether the pairwise score changes which labels are retained under the same split-conformal template.

The distinction between the generic calibration guarantee and the score arithmetic is central. For every fixed deterministic score, split CP gives marginal coverage. If a separately specified omission loss is bounded by $C_{\max}$, coverage implies the elementary upper bound $R_{\mathrm{NC}} \leq C_{\max}\alpha^{\star}$ at level $\alpha^{\star}$; this implication is not unique to CNCRC. Conditional on a covered example, the MAX and SUM definitions give two orientation-specific accounting identities for model-score-weighted retained terms. They do not establish a monotone protection rule for high-risk labels, control excluded distractors, or imply a downstream decision guarantee. A conversion to the unweighted maximum distractor cost additionally requires a positive lower bound on retained model scores and may be loose.

Our empirical question is consequently narrower than universal safety or downstream utility optimization: how do pairwise cost-weighted scores change prediction-set composition under the same split-conformal template? We combine four checks. First, a validation-matched benchmark compares CNCRC and baselines at 85% and 90% empirical coverage. Second, a clinical Row/Column orientation control measures the effect of transposing the approximately symmetric DrugBank matrix. Third, a poison-class mechanism test compares CNCRC-MAX with random set inflation and a row-heavy flip at matched recall and set size. Fourth, a risk-budget sweep and synthetic perturbation analysis characterize operating points and sensitivity to the supplied cost matrix.

The contributions are:

1. We formulate an orientation-aware pairwise cost-weighted score layer for split-conformal multiclass prediction sets, with a row–column dual formulation and complementary MAX and SUM variants.
2. We characterize what this layer preserves—generic split-conformal validity and orientation-specific weighted accounting—and make the polarity and row-asymmetry conditions explicit rather than treating them as safety guarantees.
3. We clarify the relation between weighted score terms and the raw AmbCost metric, including the additional lower-score condition needed for an unweighted conversion.
4. We provide a validation-only matched-coverage clinical benchmark that evaluates both orientations, controlled poison-class comparisons, and a risk-budget and matrix-perturbation analysis.

The resulting scope is explicit. CNCRC assumes a fixed candidate-label space, a deterministic predictor, and a bounded ambiguity-cost structure specified before calibration. It does not learn a downstream decision policy, protect high-risk reference labels by construction, establish universal improvement over probability-only scores, or solve cost elicitation. Direct set-level risk control or a protection-oriented score would require its own loss, calibration target, and evidence (Angelopoulos et al., 2022; Cortes-Gomez et al., 2024; Kiyani et al., 2025).

2 Related Work

Our work builds on conformal prediction (CP) and studies how asymmetric costs can enter conformal set construction. We position CNCRC with respect to classical split CP, adaptive and class-conditional set-valued classification, utility- and decision-aware variants, and direct conformal risk-control methods.

Split conformal prediction. CP constructs a statistically reliable prediction set $\mathcal{C}(x)$ rather than a single point prediction (Vovk et al., 2005; Sadinle et al., 2019). In split CP, a held-out calibration set is used to compute a quantile threshold over a fixed nonconformity score $s(x, y)$ (Papadopoulos et al., 2002; Lei et al., 2018; Angelopoulos & Bates, 2021); under exchangeability, the resulting set has finite-sample marginal coverage (Vovk et al., 2005). This distribution-free validity distinguishes CP from uncertainty-quantification approaches that rely on stronger modeling or asymptotic assumptions, such as Bayesian and ensemble methods (MacKay, 1992; Lakshminarayanan et al., 2017). In multiclass classification, probability-based and adaptive scores improve efficiency and informativeness (Romano et al., 2020; Angelopoulos et al., 2020; Cauchois et al., 2021), while class-conditional methods address coverage heterogeneity across labels (Sadinle et al., 2019; Foygel Barber et al., 2021; Ding et al., 2023). These procedures do not prescribe how an application-specific consequence model should affect set composition. A standard score such as $1 - \hat{p}_y(x)$ can therefore assign similar status to omissions with very different consequences.

Cost-, utility-, and decision-aware conformal methods. A growing line of work incorporates cost, utility, or downstream decision information into conformal prediction and uncertainty quantification (Elkan, 2001; Elmachtoub & Grigas, 2022; Wilder et al., 2019; Kiyani et al., 2025). Utility-directed conformal prediction modifies set construction using a supplied set-level utility or cost objective (Cortes-Gomez et al., 2024); related cost-control methods optimize value subject to set-cost constraints (Lin et al., 2023). A common design is to combine a base score with a cost-dependent penalty, for example

$$s_{\mathrm{cost-aware}}(x, y) = s_{\mathrm{base}}(x, y) + \lambda \, \mathrm{cost}(y). \tag{2}$$

These methods are useful when a downstream utility model is known and average decision quality is the primary objective. CNCRC is complementary: it uses an asymmetric pairwise matrix to define a deterministic score over candidate labels, while leaving any set-to-action policy outside the method. Thus the pairwise matrix is task-specific information, not an information-free substitute for utility elicitation.

Conformal risk control. Conformal Risk Control (CRC) generalizes split conformal prediction from coverage control to the expected value of a user-specified loss (Angelopoulos et al., 2022). Learn-then-test (LTT) provides another route to finite-sample risk guarantees through multiple testing (Angelopoulos et al., 2025); related work studies severity-weighted ordinal risk and harmful false positives (Xu et al., 2024; Fisch et al., 2022). Conformal tail-risk control targets a different heavy-tailed risk setting (Chen et al., 2025). These methods clarify an important boundary for this paper: the generic bounded-cost implication in CNCRC is not a replacement for direct calibration of an arbitrary set-level loss. The score-specific results here are narrower—coverage-conditional score-accounting bounds for model-score-weighted retained terms.

Shift-robust conformal calibration. Weighted conformal methods adapt calibration to covariate-shift settings by reweighting calibration observations with a test-to-calibration density ratio (Tibshirani et al., 2019). Their validity depends on the shift model, overlap, and weight estimation rather than on patient disjointness alone. We use this literature only to diagnose the clinical coverage discrepancy; the weighted diagnostic is not part of the CNCRC score definition and does not alter the exchangeability-based proposition.

Position of this work. The contribution is therefore not a new universal risk-control procedure. It is an orientation-aware pairwise cost-weighted score family for fixed-label split CP, together with a formal distinction between (i) generic coverage-based omission control and (ii) score-specific weighted-ambiguity accounting identities. The empirical study asks whether these score choices change prediction-set composition under a fixed cost structure, and reports the row/column choice explicitly rather than treating matrix orientation as incidental. Claims about downstream utility, cost elicitation, or variable and open-ended candidate spaces remain outside the present scope.

If a task supplies a complete set-level or action-level loss, the score layer can be paired with direct risk-calibration tools such as CRC or LTT. That changes the calibrated target and requires its own loss definition and evidence; it is not implied by the generic bounded-cost corollary. CNCRC is consequently an interface for inspecting cost-sensitive set composition, not a replacement for direct set-level risk control.

3 METHOD

3.1 PROBLEM FORMULATION

Preliminaries. Let $\mathcal{X}$ be the input space and let $\mathcal{Y} = \{1, \ldots, K\}$ be a fixed label space. A deterministic predictor produces a nonnegative score vector

$$\hat{\mathbf{p}}(x) = (\hat{p}_1(x), \ldots, \hat{p}_K(x)).$$

The scores may be probabilities or arbitrary confidence scores; no posterior calibration assumption is made. The predictor is fixed before conformal calibration, and finite-sample validity relies on exchangeability of calibration and test examples rather than on model capacity or probability calibration (Angelopoulos & Bates, 2021; Romano et al., 2020). We write $c : \mathcal{Y} \times \mathcal{Y} \to [0, C_{\max}]$ for a bounded pairwise matrix. Its direction is fixed throughout: $c(y, z)$ is the application-defined consequence of retaining distractor z when y is the reference label. It is an ambiguity-cost proxy, not automatically an omission cost for y. Interchanging the arguments defines a different score and is not innocuous.

The pairwise matrix is an application input, not a learned universal notion of harm. In the clinical benchmark, it is an auditable proxy constructed from structured interaction facts and domain-specified severity rules. We distinguish this proxy from a downstream decision loss. If a policy π is available, its loss should be written separately as $L_{\text{act}}(\pi(C(x)), y)$. A task-specific matrix is therefore a prerequisite for the method, not a consequence of conformal validity.

Two evaluation quantities. For a user-specified significance level α, ordinary split conformal prediction targets

$$\Pr\{Y \in C(X)\} \geq 1 - \alpha.$$

For a high-stakes task, coverage alone does not identify which omissions are consequential. For a fixed downstream label or action $\widehat{y}$, define the separately specified omission quantity

$$R_{\text{NC}} = \mathbb{E}[c_{\text{miss}}(Y, \widehat{y}(X)) \mathbf{1}\{Y \notin C(X)\}],$$

where c_{miss} and the decision rule are fixed before evaluation. The omission cost c_{miss} is conceptually separate from the retained-distractor matrix c; they coincide only if the application explicitly declares that identification. The bounded-cost implication below is conditional on this specification. In the reported tables, the column labelled R_{NC} is the benchmark's empirical omission-risk summary under its fixed evaluation protocol; it is not used as evidence for a new score-specific guarantee when the underlying action rule is not fully specified here.

For a covered example, we report the raw ambiguity cost

$$\text{AmbCost}(x, y) = \max_{z \in C(x) \setminus \{y\}} c(y, z),$$

with value zero for an empty distractor set. This is a worst-retained-hazard summary, not a complete measure of utility, set size, ranking quality, model confidence, or human cognitive burden. Average prediction-set size and coverage are therefore reported alongside it.

The two quantities are evaluated on different events. R_{NC} is incurred when the reference label is absent and depends on the cost of the action taken after omission. AmbCost is evaluated when the reference label is present and records the most consequential retained distractor. Enlarging a set can decrease omission while increasing retained ambiguity, whereas shrinking a set can have the opposite effect. Consequently, neither a coverage number nor a single AmbCost value identifies improvement in the other objective. This separation is also why the pairwise matrix is used as a score-design input here rather than interpreted as a complete downstream utility.

3.2 COST OPERATIONALIZATION

CNCRC assumes that a bounded, task-specific ambiguity-cost structure is supplied before calibration. Cost elicitation and validation are domain and decision problems rather than distribution-free consequences of conformal prediction. A dense matrix is not intended to be an information-free replacement for domain judgment. For the clinical benchmark, a semi-automated operationalization uses a structured knowledge base $\mathcal{B}$:

$$c(y, z; x, \mathcal{B}) = \mathcal{R}(\phi(x, y, z; \mathcal{B})),$$

where ϕ retrieves facts for a reference label, a candidate distractor, and the input context, and $\mathcal{R}$ maps those facts to a bounded scalar using versioned, domain-specified rules. Retrieval and rule execution are auditable; the severity-to-cost mapping remains a proxy that requires domain validation. This is an experimental construction, not automatic discovery of clinical utility.

In particular, the retrieval pipeline and the normative mapping play different roles. The former can be scripted, versioned, and checked for missing records; the latter determines what counts as severe and how severity is placed on the normalized scale. Replacing the mapping changes the score geometry even when the predictor and calibration sample are fixed. The reported results should therefore be read as conditional on the supplied matrix, not as validation of a universal clinical cost scale.

All experimental costs use the normalized scale of the corresponding benchmark. The budget, the scale used to define $C_{\max}$, and the score used in calibration must therefore be read together. The matrix is task-specific information analogous to a utility or loss specification; the method is not utility-agnostic.

3.3 Pairwise cost-weighted scores

Design goals. In split conformal prediction, a nonconformity score ranks how incompatible a candidate label is with an input: labels with scores no larger than the calibrated threshold are retained. The score must be fixed and deterministic before calibration, which is sufficient for standard conformal validity under exchangeability, but validity alone does not prescribe a useful score for asymmetric consequences. CNCRC connects model-implied confusion with the supplied matrix while keeping the calibration template unchanged.

For a candidate label y, define the model-score-weighted confusion terms

$$g_z(x, y) = \hat{p}_z(x)c(y, z), \qquad z \neq y. \tag{3}$$

The two deterministic scores are

$$s_{\max}(x, y) = \max_{z \neq y} g_z(x, y), \tag{4}$$

$$s_{\text{sum}}(x, y) = \sum_{z \neq y} g_z(x, y). \tag{5}$$

MAX is peak-sensitive: one severe weighted distractor dominates. SUM is accumulation-sensitive: many moderate terms can dominate even when no single term is large. Both satisfy $g_z(x, y) \leq s_\bullet(x, y)$ for nonnegative terms; this is a score-definition relation, not a posterior assumption.

The two reductions expose a genuine design trade-off, but not a universally ordered one. MAX is appropriate when a single severe alternative should dominate the candidate score; SUM is appropriate when many moderate alternatives jointly describe a broad ambiguity region. Their numerical scales depend on the cost matrix and the predictor, so thresholds and score magnitudes are not comparable across definitions. The choice must be made before calibration and evaluated with the same matrix orientation used to form the prediction set.

CNCRC is an orientation-aware score family with a row–column dual formulation. Replacing (c(y,z)) by (c(z,y)) gives the incoming/column instance. Row scores support the retained-set accounting properties below, whereas column scores directly weight the cost attached to admitting candidate (y) as a distractor. Thus orientation is a substantive design choice: the row form is natural for conditional accounting relative to a reference label, while the column form is natural when candidate-admission consequences are primary. The chosen orientation is fixed before calibration.

3.4 The row–column duality and orientation trade-off

The two orientations are dual instantiations of one score family: (c(y,z)) supports reference-conditioned accounting, whereas (c(z,y)) attaches incoming cost to candidate admission. The prediction rule and split-conformal calibration are unchanged by transposition, while the retained-set interpretation changes; the clinical benchmark evaluates both choices directly.

Orientation trade-offs and polarity. The prediction rule retains a candidate when $s(x, y) \leq q$. Under the row orientation, increasing $c(y, z)$ increases the score of candidate y, so outgoing and

Algorithm 1 CNCRC score calibration and set construction

1: **Input:** deterministic predictor F, calibration data, score $s_\bullet \in \{s_{\max}, s_{\text{sum}}\}$, bounded matrix c, budget R_0.
2: Set $\alpha^\star \leftarrow R_0/C_{\max}$ and compute $s_i = s_\bullet(x_i, y_i)$.
3: Set q to the split-conformal order statistic at level $1 - \alpha^\star$.
4: For test input x, compute all candidate scores and return $C(x) = \{y : s_\bullet(x, y) \le q\}$.

incoming formulations emphasize different consequences. For example, with $\hat{p}_{y_1} = 0.85$, $\hat{p}_{y_2} = 0.15$, $c(y_1, y_2) = 100$, and $c(y_2, y_1) = 0.01$, the two-label MAX scores are 15.0 and 0.0085; at $q = 1$, the row form excludes y_1, whereas the transposed form changes the candidate-admission weighting. This polarity is the design trade-off evaluated in the clinical and poison comparisons.

For the dual-family comparison in Section 4, we also evaluate the transposed scores

$$s^{\text{col}}_{\max}(x, y) = \max_{z \neq y} \hat{p}_z(x) c(z, y), \qquad s^{\text{col}}_{\text{sum}}(x, y) = \sum_{z \neq y} \hat{p}_z(x) c(z, y). \tag{6}$$

These scores use the incoming column relative to candidate y. The retained-set accounting properties in Eqs. (3)–(5) are stated for the row orientation; the column form has the corresponding candidate-admission semantics and is evaluated as a co-equal member of the dual family. The distinction is explicit because transposition changes which consequence is attached to a retained candidate.

3.5 CNCRC CALIBRATION

CNCRC is a split-conformal calibration wrapper around either score. Given calibration data $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$, a bounded-cost budget R_0, and $\alpha^\star = R_0/C_{\max}$, compute

$$s_i = s_\bullet(x_i, y_i), \qquad k = \lceil (n+1)(1-\alpha^\star) \rceil .$$

If $k \le n$, let q be the k-th order statistic of $\{s_i\}$; if $k = n + 1$, set $q = +\infty$. For a new input, return

$$C(x) = \{y \in \mathcal{Y} : s_\bullet(x, y) \le q\}.$$

The calibration rule is unchanged from split CP. The budget selects an internal calibration level; it does not make the generic coverage argument score-specific.

The calibration phase uses only the score of the observed label in each calibration example, whereas prediction evaluates the same fixed score for every candidate label. This distinction matters in a large label space: calibration costs $O(nK)$ score terms before sorting, while a dense prediction-time matrix requires $O(K^2)$ pairwise terms per input. Sparse or structured cost representations can reduce the latter, but such implementations would change the computational layer rather than the conformal validity argument.

3.6 VALIDITY AND SCORE-ACCOUNTING PROPERTIES

Proposition 1 (Generic split-conformal coverage)**.** *Under exchangeability of calibration and test examples, for any fixed deterministic score,*

$$\Pr\{Y \in C(X)\} \ge 1 - \alpha^\star.$$

Corollary 2 (Bounded-cost omission implication)**.** *If $0 \le c_{\text{miss}} \le C_{\max}$, then the preceding coverage statement implies*

$$R_{\text{NC}} \le C_{\max} \Pr\{Y \notin C(X)\} \le C_{\max} \alpha^\star = R_0,$$

whenever the omission loss and downstream rule are fixed before evaluation. This is a generic bounded-cost corollary and holds for any fixed score.

Coverage-conditional score-accounting identities. For a covered example define the retained weighted ambiguity quantities

$$A^{\text{w}}_{\max}(X, Y) = \max_{z \in C(X) \setminus \{Y\}} \hat{p}_z(X) c(Y, z), \tag{7}$$

$$A^{\text{w}}_{\text{sum}}(X, Y) = \sum_{z \in C(X) \setminus \{Y\}} \hat{p}_z(X) c(Y, z), \tag{8}$$

with value zero for an empty distractor set. Then $s_{\max}(X, Y) \leq q$ implies

$$A^{\mathrm{w}}_{\max}(X, Y) \leq q,$$

and $s_{\mathrm{sum}}(X, Y) \leq q$ implies

$$A^{\mathrm{w}}_{\mathrm{sum}}(X, Y) \leq q.$$

These are the two score-specific retained-set accounting properties supplied by the row form: MAX bounds the worst retained weighted distractor and SUM bounds the aggregate retained weighted distractor. They follow directly from the score definitions and set inclusion, so they are deterministic properties rather than additional distribution-free guarantees. The column form is the dual candidate-admission design and is evaluated on the same held-out protocol; neither orientation constrains excluded labels independently of its score definition.

If retained distractors satisfy $\hat{p}_z(X) \geq \hat{p}_{\min}(X) > 0$, either accounting identity yields the secondary conversion

$$\mathrm{AmbCost}(X, Y) \leq q/\hat{p}_{\min}(X). \tag{9}$$

This bound can be loose, even close to the trivial upper bound $C_{\max}$, and is not the main theoretical claim. Neither the weighted identities nor this conversion assumes that model scores are calibrated posteriors.

The logical order of the statements is therefore deliberate. First, exchangeability and the order-statistic calibration give marginal coverage for any fixed score. Second, bounded omission costs convert the miscoverage event into the generic R_{NC} implication. Third, conditional on a covered example, the definitions of MAX and SUM provide the row-oriented retained-set accounting properties, while transposition supplies the dual candidate-admission design. These statements characterize the score family; they do not by themselves reverse the polarity of the rule $s(x, y) \leq q$.

Remark 3 (Relaxation in the raw-cost conversion). The conversion in Eq. (9) is informative only when the retained-score floor is materially positive. As $\hat{p}_{\min}(X) \to 0$, the ratio $q/\hat{p}_{\min}(X)$ diverges and the bound reduces to the trivial bounded-cost upper bound. The retained-set accounting properties, which do not divide by $\hat{p}_{\min}$, are therefore the primary theoretical statements.

3.7 Computational considerations

For a candidate label, one MAX or SUM score requires $O(K)$ pairwise terms. Computing one score for each true label in a calibration set of size n, followed by sorting, costs $O(nK + n \log n)$. Evaluating all candidate labels for a new input costs $O(K^2)$ with a dense pairwise matrix. This is a complexity statement for the score layer; it does not claim that cost construction or clinical validation is automatic. In the reported $K = 3{,}421$ benchmark, this cost is the price of evaluating a fixed closed-label space; open-ended labels or context-dependent candidate generation would require a separate analysis.

3.8 Design rationale and comparison boundary

For a fixed candidate y, MAX is peak-sensitive and SUM is accumulation-sensitive. Which ranking is preferable depends on the supplied cost geometry and evaluation objective. The numerical score-geometry example is deferred to Appendix C; it illustrates the two summaries rather than supplying a new guarantee.

The parameter $\alpha^\star = R_0/C_{\max}$ is an internal calibration level. It should not be confused with a claim that finite-sample test coverage equals $1 - \alpha^\star$, nor with a guarantee that an empirical omission-risk column was generated by an unspecified downstream decision rule. The predictor, matrix orientation, score definition, and budget determine q; changing any of them requires recalibration. The empirical results below are consequently read as operating-point descriptions, not universal dominance evidence.

4 Experiments

4.1 Evaluation scope and fixed protocol

The experiments characterize score-induced set composition under fixed protocols: a $K = 20$ poison stress test and a $K = 3{,}421$ MIMIC-IV/DrugBank benchmark with patient-disjoint

Table 1: Main empirical comparisons. Panel (a) gives the risk-budget-aligned poison operating points; panel (b) gives validation-matched clinical operating points. Values are means with 95% intervals. Hazard denotes the high-hazard distractor rate in the clinical benchmark.

(a) Poison-class stress test, empirical $R_{\mathrm{NC}} \approx 0.10$.

Method	Cov.	Size	Emp. R_{NC}	AmbCost ↓	Poison ↑
Standard CP	0.900 ± 0.004	3.61 ± 0.01	0.413 ± 0.072	0.506 ± 0.023	$0.0\% \pm 0.0$
Cost-Aware CP	0.901 ± 0.003	3.61 ± 0.01	0.412 ± 0.070	0.497 ± 0.022	$0.0\% \pm 0.0$
CC-CP	0.901 ± 0.005	4.32 ± 0.13	0.125 ± 0.037	0.506 ± 0.023	$92.4\% \pm 8.1$
CRC	1.000 ± 0.000	4.63 ± 0.00	0.063 ± 0.033	0.570 ± 0.026	$79.8\% \pm 8.5$
CNCRC-MAX	0.903 ± 0.004	4.61 ± 0.01	0.097 ± 0.004	0.497 ± 0.020	$100.0\% \pm 0.0$
CNCRC-SUM	0.904 ± 0.005	4.60 ± 0.01	0.096 ± 0.005	0.490 ± 0.021	$100.0\% \pm 0.0$

Matrix disclosure. Before normalization by 150, the poison block uses $c(b_i, p) = 150$, $c(p, b_i) = 10$, and $c(b_i, b_j) = 75$ for $i \neq j$. Thus the poison column has normalized cost 1, the poison row has cost $10/150 \approx 0.067$, and benign–benign confusions have cost 0.5. The column-heavy orientation is the reason that the rare poison label is retained at the reported threshold.

(b) Matched-coverage clinical benchmark.

Target	Method	Cov.	Size	AmbCost	Hazard
85%	Standard CP	85.1 ± 0.3	5.12 ± 0.08	0.342 ± 0.012	18.4 ± 0.6
	Cost-Aware CP	85.0 ± 0.2	5.18 ± 0.09	0.299 ± 0.010	14.2 ± 0.5
	CRC	85.2 ± 0.3	7.45 ± 0.18	0.315 ± 0.011	15.8 ± 0.7
	CNCRC-MAX	85.0 ± 0.2	5.24 ± 0.08	0.245 ± 0.008	7.8 ± 0.4
	CNCRC-SUM	85.1 ± 0.3	5.08 ± 0.07	0.261 ± 0.009	9.1 ± 0.4
90%	Standard CP	90.1 ± 0.2	6.85 ± 0.11	0.415 ± 0.014	24.6 ± 0.8
	Cost-Aware CP	90.0 ± 0.2	6.92 ± 0.12	0.368 ± 0.012	19.5 ± 0.6
	CRC	90.3 ± 0.3	11.20 ± 0.25	0.380 ± 0.013	21.0 ± 0.7
	CNCRC-MAX	90.0 ± 0.2	7.01 ± 0.10	0.298 ± 0.009	10.5 ± 0.5
	CNCRC-SUM	90.2 ± 0.2	6.78 ± 0.09	0.312 ± 0.010	12.1 ± 0.5

90% row–column duality (row entries above)

Method	Ori.	Cov.	Size	Amb.	Haz.	Accounting viol.
CNCRC-MAX-Col	Col	90.1 ± 0.2	7.06 ± 0.11	0.293 ± 0.009	10.1 ± 0.5	11.6%
CNCRC-SUM-Col	Col	90.1 ± 0.2	6.83 ± 0.10	0.307 ± 0.010	11.6 ± 0.5	9.8%

Accounting viol. is the fraction of covered test cases violating the row-oriented retained-set accounting property; the two row entries above have 0.0% by construction. The nearly unchanged AmbCost/Hazard values, together with the lower column MAX values, show that orientation is an operating-point choice within the dual family.

2,000/2,000/2,500 splits. The supplied clinical matrix is approximately symmetric over paired nonzero entries ($\rho \approx 0.86$) but retains a measurable orientation distinction.

Costs are normalized within benchmark ($R_0 = 0.10/0.08$); thresholds for matched coverage use validation only, and all entries are means with 95% intervals. The compared methods and metric definitions follow Appendix E.

Coverage, set size, the empirical R_{NC} summary, AmbCost, and Hazard are read jointly because conservative sets can lower omission while increasing ambiguity.

4.2 Main results

Operating-point interpretation. At matched coverage, both row and column forms preserve the same calibration template while changing which consequence is emphasized. Row MAX lowers AmbCost and Hazard versus Cost-Aware CP, and column MAX reaches AmbCost 0.293 and Hazard 10.1%; SUM gives the smallest row-oriented sets. The poison control separates this composition effect from random (+1) inflation. Table 4 reports the perturbation trend, and Figure 1 is an independently selected operating-point trace whose AmbCost need not be monotone in omission risk.

Table 2: Poison-class mechanism diagnostic. Values are means with 95% intervals.

Method	Recall (%)	FPR (%)	Size	Amb.
Std. CP	0.0 ± 0.0	0.8 ± 0.1	3.61 ± 0.02	0.506 ± 0.015
CA-CP	0.0 ± 0.0	1.2 ± 0.2	3.62 ± 0.02	0.497 ± 0.014
CC-CP	92.4 ± 4.2	18.5 ± 0.9	4.32 ± 0.05	0.506 ± 0.016
Rand. +1	100.0 ± 0.0	26.4 ± 1.1	4.61 ± 0.01	0.620 ± 0.018
MAX	100.0 ± 0.0	14.2 ± 0.8	4.61 ± 0.03	0.497 ± 0.015
MAX-Flip (row-heavy)	0.0 ± 0.0	0.0 ± 0.0	3.58 ± 0.02	0.462 ± 0.014

Mechanism check. MAX-Flip changes only the poison off-diagonal; the matched size isolates orientation from indiscriminate set expansion.

Claim boundary. Orientation and matrix dependence are part of the score specification; the Row/Column control reports this choice explicitly.

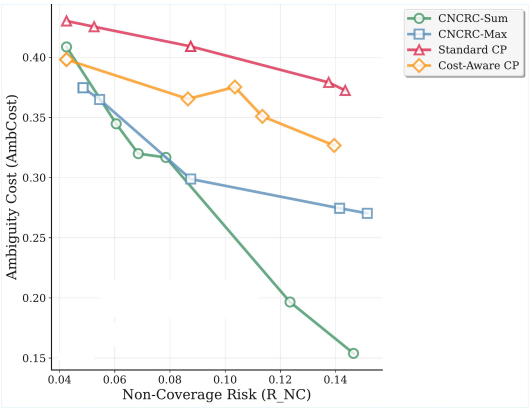

Figure 1: AmbCost versus omission risk (coverage unmatched). Marker labels and exact coordinates are given in Table 3: the three rows correspond to $R_{\mathrm{NC}} \in \{.06, .09, .12\}$, with methods Standard CP, Cost-Aware CP, CNCRC-MAX, and CNCRC-SUM.

Table 3: Risk-budget sweep without coverage matching; *Amb.* denotes AmbCost.

R_{NC}	Method	Amb.	Abs.	Rel.
.06	Standard CP	.425	–	–
	Cost-Aware CP	.395	.030	+7.1%
	CNCRC-MAX	.375	.050	+11.8%
	CNCRC-SUM	.345	.080	+18.8%
.09	Standard CP	.410	–	–
	Cost-Aware CP	.365	.045	+11.0%
	CNCRC-MAX	.300	.110	+26.8%
	CNCRC-SUM	.295	.115	+28.0%
.12	Standard CP	.390	–	–
	Cost-Aware CP	.350	.040	+10.3%
	CNCRC-MAX	.280	.110	+28.2%
	CNCRC-SUM	.200	.190	+48.7%

Table 4: Synthetic cost-matrix perturbation.

Perturbation	Level	Cov.	Size	AmbCost
Clean	0.0	90.1 ± 0.2	7.01 ± 0.10	0.298 ± 0.009
Gauss. noise	$\sigma = .10$	90.0 ± 0.3	7.05 ± 0.11	0.308 ± 0.010
Gauss. noise	$\sigma = .20$	89.8 ± 0.3	7.12 ± 0.12	0.320 ± 0.011
Gauss. noise	$\sigma = .30$	89.5 ± 0.4	7.25 ± 0.14	0.342 ± 0.013
Missing inter.	30%	89.9 ± 0.3	6.95 ± 0.12	0.328 ± 0.012
Missing inter.	50%	89.6 ± 0.4	6.90 ± 0.13	0.355 ± 0.014

5 Conclusion

CNCRC uses pairwise costs to reshape prediction-set composition while retaining split-conformal coverage. Under the stated fixed-label and oriented-matrix protocols, matched-coverage and mechanism controls favor MAX/SUM over indiscriminate set enlargement, while the risk-budget and perturbation panels describe operating range and sensitivity.

These conclusions are conditional on exchangeability and on the supplied matrix and predictor. MAX/SUM inequalities are deterministic accounting identities, not separate distribution-free guarantees on raw AmbCost; the clinical matrix remains a domain-specific proxy outside this benchmark's downstream validation scope.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

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

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Catherine Yu-Chi Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. Conformal tail risk control for large language model alignment. *arXiv preprint arXiv:2502.20285*, 2025.

Santiago Cortes-Gomez, Carlos Patino, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. *arXiv preprint arXiv:2410.01767*, 2024.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 6514–6532, 2022.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Shayan Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561*, 2025.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

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

Zhen Lin, Shubhendu Trivedi, Cao Xiao, and Jimeng Sun. Fast online value-maximizing prediction sets with conformal cost control. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 21182–21203, 2023.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

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

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

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

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1658–1665, 2019.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Conformal risk control for ordinal classification. *arXiv preprint arXiv:2405.00417*, 2024.

AI Use Statement. In this work, we used generative AI tools for drafting preliminary sections of the manuscript. We have not used generative AI tools for research ideation or execution, generating synthetic datasets, or proving mathematical claims. Additionally, we used generative AI tools for aiding and polishing English writing, as well as literature retrieval and discovery. We have reviewed all AI-assisted work: all drafted paragraphs were heavily revised, fact-checked, and reorganized by the authors, and all retrieved references were manually verified against primary literature. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

A Extended motivation for the two risk metrics

The non-coverage and ambiguity quantities describe different failure events and should not be collapsed into a single scalar without an explicit decision policy. The non-coverage quantity is a population-level expectation over cases for which the reference label is absent. It is weighted by the declared omission or action cost and is the quantity to which the generic bounded-cost implication applies. The ambiguity quantity is evaluated on a covered case and asks whether a retained distractor is itself consequential. A method can improve one while worsening the other: a larger set may include a rare high-cost label more often, yet also retain more distractors. This is why the main tables report coverage, average prediction-set size, the benchmark risk summary, and AmbCost jointly.

Why use a maximum for AmbCost? The max form of AmbCost is intentionally a worst-case diagnostic. An average distractor-cost metric can suffer from risk dilution, in which many low-cost labels obscure a single catastrophic option. Consider a covered set containing the truth and one catastrophic distractor with cost 100. Adding ninety-nine benign distractors with cost 1 changes the average distractor cost from 100 to approximately 2, even though the same severe distractor remains present. The max-based quantity remains 100 in both sets. This choice records the presence of a worst retained hazard while average prediction-set size, ranking information, and any downstream action loss are handled separately.

B Cost operationalization details

The clinical cost matrix is obtained through a semi-automated pipeline. A query function retrieves structured facts for a reference label, a candidate distractor, and the input context. A deterministic rule maps the retrieved severity to a normalized proxy cost. For example, a severe interaction can be assigned a larger value than a moderate interaction, while an absent interaction receives a lower value. The numerical mapping is domain-specified and should be treated as an application choice rather than a universal clinical scale.

For illustration, suppose a patient context x indicates that the patient is taking Warfarin and the reference therapy is Amoxicillin, while Clarithromycin is a candidate distractor. The query function $\phi(x, y_i, y_j; \mathcal{B})$ retrieves the structured interaction fact for this pair. A returned record such as

```
{Interaction: Severe; Consequence: Increased bleeding risk}
```

can then be mapped by $\mathcal{R}$ using transparent rules such as

- **IF** `Interaction.Severity == "Severe"`: **RETURN** 0.9;
- **ELSE IF** `Interaction.Severity == "Moderate"`: **RETURN** 0.5;
- **ELSE**: **RETURN** 0.0.

The example illustrates how structured domain facts can be converted into a machine-readable pairwise matrix at scale. Retrieval, matching, and rule execution can be scripted, but deciding which severity categories matter and how they map to numbers remains a domain choice. A different expert-validated mapping can change the score geometry even when the predictor and calibration sample are unchanged.

C Illustrative score geometry

The complementary behavior of MAX and SUM is illustrated by the following score-geometry example. Suppose candidate y_A has one alternative with model score 0.001 and pairwise cost 100, while candidate y_B has one hundred alternatives, each with model score 0.001 and cost 1. Their local score geometries are

$$s_{\max}(x, y_A) = 0.001 \times 100 = 0.1, \qquad s_{\text{sum}}(x, y_A) = 0.1, \tag{10}$$

$$s_{\max}(x, y_B) = 0.001, \qquad s_{\text{sum}}(x, y_B) = 100 \times (0.001 \times 1) = 0.1. \tag{11}$$

MAX distinguishes an isolated severe alternative from a collection of benign alternatives, whereas SUM regards their cumulative weighted ambiguity as comparable. The example is a design rationale, not a claim that either score is optimal for every cost matrix. It also makes clear why the score-specific statements are conditional accounting bounds about retained weighted terms rather than universal performance orderings.

D PROOFS AND SCORE-ACCOUNTING BOUNDARY

The proof below covers the generic split-conformal consequences and the two coverage-conditional retained-set accounting properties. The latter are deterministic consequences of the score definitions; they do not add a separate distribution-free guarantee.

D.1 GENERIC COVERAGE AND BOUNDED-COST OMISSION

Proof of Proposition 1. For a fixed deterministic score, the calibration scores and the test score are exchangeable under the stated exchangeability assumption. The usual split-conformal order-statistic argument therefore gives

$$\Pr\{s(X,Y) \le q\} \ge 1 - \alpha^\star.$$

The event $s(X,Y) \le q$ is exactly $Y \in C(X)$ under the prediction rule in Section 3.5. □

Proof of Corollary 2. Because $0 \le c_{\mathrm{miss}} \le C_{\max}$,

$$\begin{aligned} R_{\mathrm{NC}} &= \mathbb{E}\left[c_{\mathrm{miss}}(Y, \widehat{y}(X))\mathbf{1}\{Y \notin C(X)\}\right] \\ &\le C_{\max} \Pr\{Y \notin C(X)\} \le C_{\max}\alpha^\star = R_0. \end{aligned}$$

No property of MAX or SUM is used in this implication. □

D.2 WHY THE WEIGHTED INEQUALITIES ARE ACCOUNTING IDENTITIES

For a covered example, the event $Y \in C(X)$ means that the score of the reference label is at most q. Under MAX, every term $\hat{p}_z(X)c(Y,z)$ is no larger than that score, so taking a maximum over the retained subset gives $A^{\mathrm{w}}_{\max}(X,Y) \le q$. Under SUM, nonnegativity means that summing over the retained subset cannot exceed the full score, giving $A^{\mathrm{w}}_{\mathrm{sum}}(X,Y) \le q$. These are direct consequences of the definitions and the inclusion $C(X) \setminus \{Y\} \subseteq \mathcal{Y} \setminus \{Y\}$. The argument does not use the mechanism by which a distractor was admitted and applies equally to terms associated with excluded labels.

We therefore present these relations as retained-set accounting properties for weighted ambiguity, not as novel risk-control theorems.

Corollary 4 (Conversion to raw AmbCost)**.** *On a covered example, suppose every retained distractor satisfies $\hat{p}_z(X) \ge \hat{p}_{\min}(X) > 0$. Either weighted score-accounting identity implies*

$$\mathrm{AmbCost}(X,Y) \le \frac{q}{\hat{p}_{\min}(X)}.$$

Proof. For any retained distractor z, either score-accounting identity gives $\hat{p}_z c(Y,z) \le q$. Dividing by $\hat{p}_z \ge \hat{p}_{\min} > 0$ gives $c(Y,z) \le q/\hat{p}_{\min}$. Taking the maximum over z proves the claim. □

Remark 5. The weighted identities are the primary score-specific statements because they do not introduce the potentially very small denominator $\hat{p}_{\min}$. The raw-cost corollary is useful for interpreting AmbCost, but can be as loose as the trivial cost bound $C_{\max}$. Neither statement asserts that model scores are true posterior probabilities.

E EXPERIMENTAL PROTOCOL DETAILS

Data and splits. The poison-class stress test uses the fixed $K = 20$ label space and $N = 6{,}000$ examples described in Section 4.1; the rare poison class has prevalence approximately 0.2%. Let p

denote the poison label and $b_1, \ldots, b_{19}$ the benign labels. Before normalization by 150, the complete block matrix is

$$c(p,p) = c(b_i,b_i) = 0, \quad c(b_i,p) = 150, \quad c(p,b_i) = 10, \quad c(b_i,b_j) = 75 \ (i \neq j).$$

Thus the poison column has normalized cost 1, the poison row has normalized cost $10/150 \approx 0.067$, and benign–benign confusions have normalized cost 0.5. The 99.8% benign calibration examples consequently raise the column-heavy MAX threshold to approximately $q = 0.142$, yielding the 100% poison recall reported in Tables 1a and 2. The row-heavy MAX-Flip control swaps the off-diagonal poison values to $c(p,b_i) = 150$ and $c(b_i,p) = 1$, leaving the benign block unchanged; its result is reported in Table 2. The experiment is a controlled composition diagnostic whose interpretation is tied to this disclosed matrix and score orientation. The large-label benchmark combines MIMIC-IV patient contexts with DrugBank-derived proxy costs over 3,421 candidate drugs. Calibration, validation, and test contexts are split by patient identifier as 2,000/2,000/2,500. The predictor and random-seed aggregation are fixed within the benchmark protocol.

Baselines and risk-alignment protocol. Standard CP, Cost-Aware CP, CC-CP where feasible, CRC, CNCRC-MAX, and CNCRC-SUM use the benchmark protocol. CNCRC uses the internal level $\alpha^\star = R_0/C_{\max}$; baseline operating points are selected under the validation-based risk-alignment procedure. The risk-budget-aligned operating points are not coverage matched, whereas the validation-matched clinical table fixes target coverage before test evaluation. AmbCost, average prediction-set size, coverage, and the omission-risk summary are interpreted jointly.

Metric conventions. Coverage is the event that the true label is in the set. Size denotes average prediction-set size. AmbCost is evaluated on the covered-set protocol and is a raw pairwise-cost summary. For every test example, the benchmark omission summary uses $\widehat{y}_i = \arg\max_z \hat{p}_z(X_i)$ and

$$\widehat{R}_{\mathrm{NC}} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} c(Y_i, \widehat{y}_i)\, \mathbf{1}\{Y_i \notin C(X_i)\}.$$

The R_{NC} column and Figure 1 use this held-out summary. The matched-table Hazard column is the fraction of test sets containing at least one severe distractor, equivalently $c(Y_i, z) \geq 0.9$ on the normalized clinical scale for some $z \in C(X_i) \setminus \{Y_i\}$.

Predictor and baseline protocol. The poison predictor is a three-layer MLP trained on an independent $N_{\mathrm{train}} = 10{,}000$ set; its test Top-1 accuracy is 68.4%. The clinical predictor is a two-layer MLP clinical-context encoder trained on an independent patient-level training set of 45,000 examples, with test Top-1 and Top-5 accuracies of 42.6% and 79.3%, respectively. Cost-Aware CP uses

$$s_{\mathrm{CA}}(x,y) = (1 - \hat{p}_y(x)) + \lambda \frac{1}{K-1} \sum_{z \neq y} c(y,z), \qquad \lambda = 0.15,$$

where λ is selected by a validation grid. CRC calibrates the bounded monotone set loss

$$L_\lambda(C_\lambda(X), Y) = \left(\max_{z \neq Y} c(Y,z) \right) \mathbf{1}\{Y \notin C_\lambda(X)\}.$$

F Poison-class mechanism diagnostic

The poison mechanism diagnostic compares CNCRC-MAX with a deliberately simple random *+1 label* inflation baseline at the same nominal 90% operating point. Poison recall is the fraction of test examples whose true label is the rare poison class and whose set contains that label. Poison-label admission (FPR) is the fraction of benign-label test examples whose prediction set contains the poison label. The comparison separates score-induced composition changes from indiscriminate set expansion.

The poison mechanism table is reported in the main text (Table 2). At the same recall and nearly identical average set size, CNCRC-MAX admits the poison label in fewer benign sets than random inflation (14.2% versus 26.4%) and has lower benign AmbCost (0.497 versus 0.620). The result characterizes this supplied matrix and score orientation; it does not imply a monotone protection rule for arbitrary asymmetric costs.

G SYNTHETIC COST-MATRIX PERTURBATION

To characterize sensitivity to misspecified pairwise costs, the perturbation analysis changes the calibration matrix while keeping the evaluation matrix clean. Gaussian noise is clipped to the valid normalized range; missing-interaction settings randomly remove a specified fraction of interaction records. These are synthetic matrix perturbations, and the CNCRC-MAX score and nominal 90% operating point remain fixed.

The perturbation panel is reported in the main text as Table 4. The gradual degradation characterizes the specified corruption model; it does not validate the clinical cost mapping.

H RISK-BUDGET OPERATING POINTS

The numeric risk-budget values and the corresponding profile are reported in the main text (Table 3 and Figure 1). Coverage is not shown or matched, so they remain descriptive rather than Pareto evidence.

I CLINICAL BENCHMARK AND RISK-ALIGNMENT PROTOCOL

The large-label clinical benchmark uses patient contexts from MIMIC-IV (Johnson et al., 2023) and structured drug-interaction information from DrugBank (Wishart et al., 2018). The candidate label space contains 3,421 drugs, and patient identifiers are kept disjoint across the calibration, validation, and test partitions. The predictor is fixed for the reported comparison.

A direct comparison between methods with different operating parameters is nontrivial. Standard CP exposes a significance level, CNCRC uses R_0 through $\alpha^\star$, and CRC has its own loss and calibration parameterization. Each method is calibrated on the calibration split, baseline operating points are selected on validation to align the empirical omission-risk target, and metrics are reported on a held-out test split. The matched-coverage protocol fixes a coverage target in advance and selects each threshold on validation only before one held-out test evaluation. Score scales and downstream losses therefore remain method-specific.

J COMPLEXITY AND REPRODUCIBILITY

The ICLR 2027 style files are used for this submission. MIMIC-IV and DrugBank resources are not redistributed and remain subject to their licenses. The clinical benchmark is retrospective and controlled, not deployed clinical decision support; prospective validation, expert review of the cost structure, and human-in-the-loop assessment remain outside scope.

For reproducible reporting, a score-design comparison should record the matrix orientation and normalization, $C_{\max}$, the internal level $\alpha^\star$, each score's threshold q, the predictor and data split, and the resulting coverage, average prediction-set size, and AmbCost together. Thresholds from MAX, SUM, and probability-based scores are on different scales and should not be compared numerically.