Severity-Controlled Prediction Sets for Medication Recommendation

Yu Gu[1*] **Zijun Yu**[1*] **Chi-Kuang Yeh**[2] **Xinyu Wang**[1] **Ziyang Song**[3†]
[1]McGill University, [2]Georgia State University, [3]Ohio University

Abstract

Medication recommendation is a multi-label prediction task. However, existing prediction-set evaluation and calibration methods rely mainly on exact-overlap metrics and therefore treat all missed medications as equally severe, overlooking how closely the prediction set matches them within the medication hierarchy. To address this, we introduce **Conformal Severity Control (CSC)**, a post-hoc framework for calibrating medication prediction sets under hierarchy-aware mismatch severity. CSC uses the Anatomical Therapeutic Chemical (ATC) hierarchy to quantify mismatch severity and controls patient-level severity without retraining the predictor, supporting both average severity and upper-quantile severity for the highest-severity mismatches. Experiments on MIMIC-IV show that CSC tracks user-specified severity tolerances and shifts prediction sets away from distant mismatches toward exact or close matches.

1 Introduction

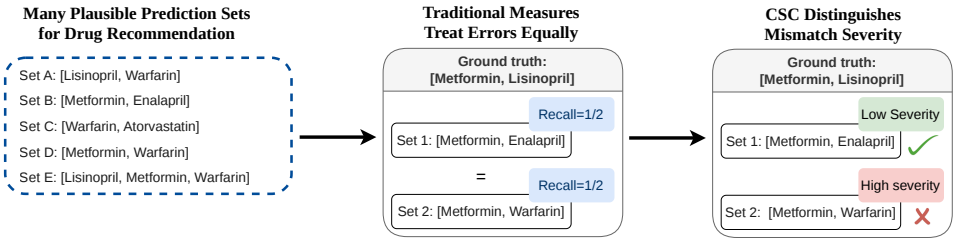


Figure 1: **Motivation for Conformal Severity Control (CSC)**. Prediction sets can achieve the same recall while differing substantially in ATC mismatch severity. CSC selects a score threshold on calibration data so that the expected severity stays below a user-specified tolerance.

Modern medication recommendation systems increasingly use deep learning models and LLM-based predictors to recommend drugs from electronic health records (EHRs) [Ali et al., 2023, Lv et al., 2026, Wang et al., 2026]. These models assign scores over large drug vocabularies, which can be used to form candidate medication sets because multiple drugs may be plausible for the same patient [Kompa et al., 2021, Begoli et al., 2019]. However, the evaluation and calibration of these sets largely rely on exact-overlap criteria including recall, F1, Jaccard, and false-negative rate [Shang et al., 2019, Yang et al., 2021, Ali et al., 2023]. These criteria treat every missed medication as the same error, even though a prediction set containing a therapeutically closer medication should incur lower severity than one containing only distant medications. Consequently, two prediction sets can receive the same

*Equal contribution.
†Corresponding author: ziyangs@ohio.edu.
Code is available at https://github.com/gggggyyyyyy/conformal-severity-control.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

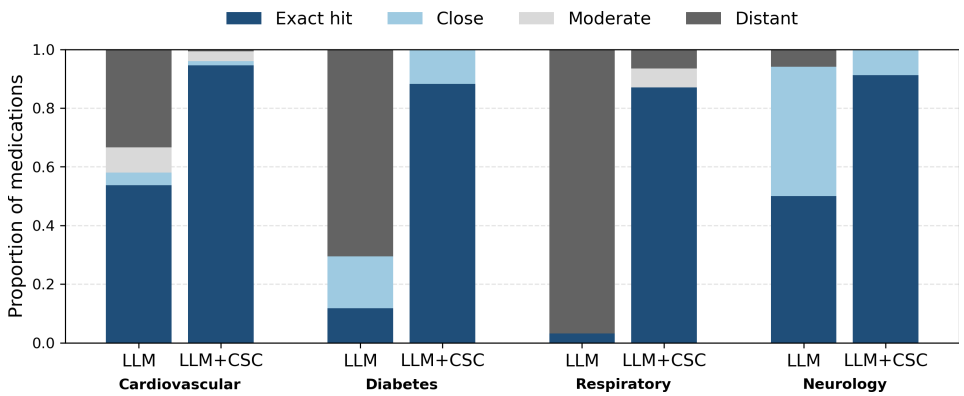


Figure 2: **Closest ATC match for disease-associated medications.** For four disease cohorts, each selected ground-truth medication is assigned its closest predicted medication and categorized as exact, close, moderate, or distant. LLM sets threshold the predicted probabilities at 0.25; CSC sets use CSC-Avg with $\alpha = 0.125$, whose cutoff is lower. The shift toward exact and close matches therefore reflects set expansion. Details are in Appendix D.1.

overlap score while differing substantially in mismatch severity (Figure 1). This motivates calibrating prediction sets to control hierarchy-aware mismatch severity.

To characterize mismatch severity in medication prediction, we analyze LLM-generated prediction sets across cardiovascular, diabetes, respiratory, and neurology cohorts. For each cohort, we focus on selected treatment medications and identify the closest predicted match under the medication hierarchy, categorizing each match as exact, close, moderate, or distant. As shown in Figure 2, missed medications in LLM-generated sets are more often matched only to distant medication predictions. In contrast, the CSC-calibrated sets reduce these distant mismatches and increase exact or close matches. This motivates post-hoc calibration under hierarchy-aware mismatch severity.

Risk-control frameworks calibrate a threshold on top of black-box prediction scores so that the expected loss of future prediction sets is bounded [Angelopoulos and Bates, 2023]. Existing frameworks such as Risk-Controlling Prediction Sets (RCPS) and Conformal Risk Control (CRC) provide post-hoc, distribution-free calibration for black-box predictors [Bates et al., 2021, Angelopoulos et al., 2024]. However, their guarantees are only as informative as the loss being controlled. When instantiated with overlap-based losses, these frameworks still treat all missed medications equally and therefore cannot distinguish therapeutically closer predictions from distant mismatches. Even after replacing binary overlap with hierarchy-aware mismatch severity, controlling only the average severity can be insufficient. For patients with multiple medications, average severity can obscure a small number of high-severity mismatches when most medications are matched exactly or closely [Zhao et al., 2024]. These limitations call for prediction set calibration that controls mismatch severity and emphasizes the most severe mismatches within each patient.

To address this gap, we introduce **Conformal Severity Control (CSC)**, a post-hoc, model-agnostic framework for calibrating medication prediction sets under hierarchy-aware mismatch severity. CSC uses the Anatomical Therapeutic Chemical (ATC) hierarchy [WHOCC, 2026] to measure how severe each medication mismatch is and calibrates prediction sets to control a user-specified patient-level severity tolerance, without retraining or modifying the underlying predictor. We further extend CSC from average severity control to upper-quantile severity control [Ben-Tal and Teboulle, 2007, Huang et al., 2026], which emphasizes the highest-severity mismatches within each patient. On MIMIC-IV, CSC tracks target severity tolerances, induces patient-specific prediction-set sizes under one global threshold, and reduces distant ATC-branch mismatches.

2 Related Work

2.1 Clinical Prediction

Prior work in clinical prediction has mainly focused on improving predictive models for specific downstream tasks [Song et al., 2024, 2026a]. In medication recommendation, models have improved

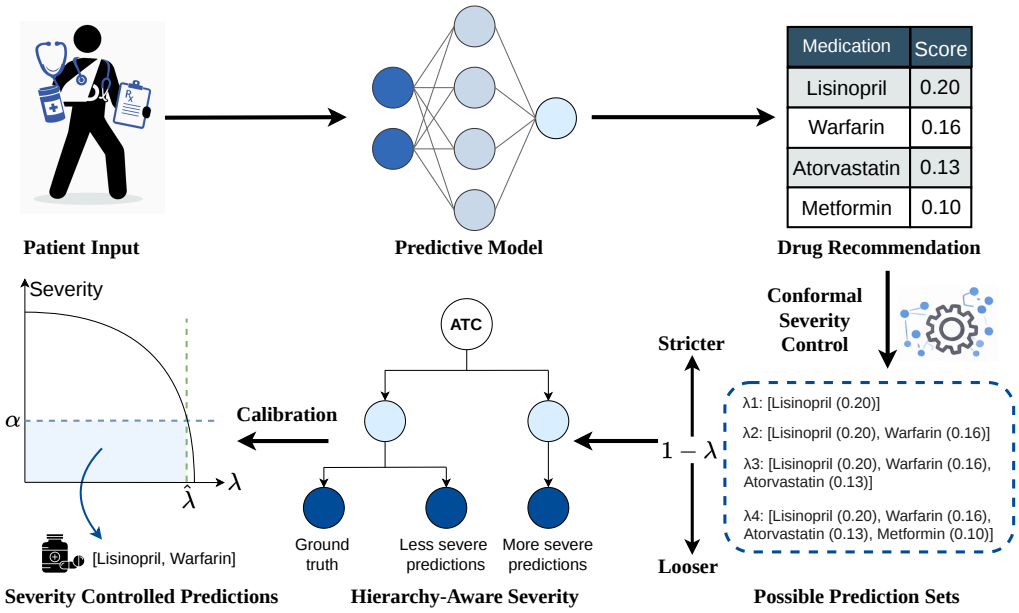


Figure 3: Overview of CSC. Given patient input, a predictive model assigns scores to candidate medications, inducing a nested family of prediction sets as the threshold λ varies. CSC evaluates these candidate sets using an ATC-based hierarchy-aware mismatch severity and calibrates λ to control the expected upper-tail patient-level severity.

candidate drug sets through stronger EHR modeling [Ali et al., 2023], safety-aware objectives [Shang et al., 2019], and richer sources of clinical information [Tan et al., 2022]. A complementary line of work incorporates medical ontologies and structured clinical relationships into the predictive model to improve predictions [Yao et al., 2023, Sun et al., 2023, Tan et al., 2024]. While these approaches improve predictive performance, they are typically specific to individual model architectures. In contrast, our method is post-hoc and model-agnostic, and can be applied to any predictive model without redesigning the predictor itself.

2.2 Clinical Calibration via Risk Control

Conformal prediction and risk-control frameworks provide distribution-free procedures for selecting prediction sets whose expected risk is controlled under exchangeability [Bates et al., 2021, Angelopoulos et al., 2024]. These frameworks are well suited to clinical prediction because they calibrate prediction sets post hoc from the scores of an already trained predictor, without requiring changes to the predictive model itself [Vazquez and Facelli, 2022]. Existing conformal methods have mainly controlled coverage-based losses, such as miscoverage, false negatives, or false positives [Fisch et al., 2022]. However, such losses do not distinguish therapeutically closer medication predictions from distant errors. We bridge this gap by quantifying the severity of each medication mismatch and calibrating prediction sets under both average and upper-quantile severity.

3 Methodology

3.1 Problem Setup

We study multi-label drug recommendation from EHR data, which provide longitudinal patient information for clinical decision support. For each patient, let $X \in \mathcal{X}$ denote the patient input and let $Y \subseteq \mathcal{K}$ denote the ground-truth medication set, where $\mathcal{K}$ is the medication label space. Each medication is represented by an ATC code, which organizes medications hierarchically from broad anatomical and therapeutic groups to finer pharmacological and chemical categories [WHOCC, 2026]. Details of the ATC taxonomy are provided in Appendix C.1.

As illustrated in Figure 3, a trained predictive model assigns a score $f_k(X) \in [0,1]$ to each medication label $k \in \mathcal{K}$. For a threshold $\lambda \in [0,1]$, we define the induced prediction set as

$$C_\lambda(X) := \{k \in \mathcal{K} : f_k(X) \geq 1 - \lambda\}, \tag{1}$$

where λ controls the inclusiveness of the prediction set, with smaller values yielding more selective sets and larger values yielding more inclusive sets.

3.2 Hierarchy-Aware Mismatch Severity

Given the prediction set $C_\lambda(X)$, we quantify how each ground-truth medication is matched by the prediction set. We view the ATC hierarchy as a rooted tree of maximum depth D, where medication labels are leaves and internal nodes represent broader medication classes. For a ground-truth medication $y \in Y$ and a predicted medication $k \in \mathcal{K}$, let $\mathrm{LCA}(y,k)$ denote their lowest common ancestor (LCA), i.e., the most specific ATC class containing both medications. If the shared ancestor is deeper in the tree, then k is closer to y in the medication taxonomy and should receive lower mismatch severity; if the shared ancestor is closer to the root, then the mismatch is more distant and should receive higher severity. We define the pairwise mismatch severity between y and k as

$$s(y,k) := 1 - \frac{\mathrm{depth}(\mathrm{LCA}(y,k))}{D}. \tag{2}$$

where $\mathrm{depth}(\cdot)$ denotes the depth of a node in the ATC tree. Thus, $s(y,k) = 0$ when y and k share the same medication label, while larger values indicate more distant medication classes. Prior work has used ATC distance to quantify therapeutic drug similarity [Wang et al., 2013] and ATC proximity to characterize therapeutic relatedness [Bohn et al., 2025].

For a ground-truth medication $y \in Y$, we define its label-level mismatch severity with respect to the prediction set $C_\lambda(X)$ by the closest predicted medication:

$$s\big(y, C_\lambda(X)\big) := \begin{cases} \min_{k \in C_\lambda(X)} s(y,k), & C_\lambda(X) \neq \emptyset, \\ 1, & C_\lambda(X) = \emptyset. \end{cases} \tag{3}$$

It assigns maximal severity when the prediction set is empty, preserving boundedness and monotonicity. The quantity $s(y, C_\lambda(X))$ is zero when $y \in C_\lambda(X)$, small when the prediction set contains a medication close to y in the ATC hierarchy, and large when all predicted medications are distant from y. Since $C_\lambda(X)$ expands as λ increases, $s(y, C_\lambda(X))$ is non-increasing in λ for every $y \in Y$.

3.3 Average and Upper-Quantile Severity Losses

The label-level mismatch severity $s(y, C_\lambda(X))$ scores a single ground-truth medication. Since a patient may have multiple ground-truth medications, we aggregate label-level severities into a patient-level loss. For $Y \neq \emptyset$, the average severity loss is

$$\mathcal{S}_{\mathrm{avg}}(Y, C_\lambda(X)) := \frac{1}{|Y|} \sum_{y \in Y} s\big(y, C_\lambda(X)\big), \tag{4}$$

and for $Y = \emptyset$ we set this loss, and the upper-quantile loss below, to zero.

Average severity can mask a few high-severity mismatches when most ground-truth medications are matched exactly or closely. To emphasize such mismatches, we control the upper tail of the within-patient severities. For $Y \neq \emptyset$, let $m = |Y|$, write $z_y = s(y, C_\lambda(X))$, and fix a quantile level $\beta \in [0,1)$. The upper-quantile severity loss is the conditional value-at-risk (CVaR) at level β of the uniform distribution over $\{z_y\}_{y \in Y}$ [Rockafellar and Uryasev, 2000, Ben-Tal and Teboulle, 2007]:

$$\mathcal{S}_\beta(Y, C_\lambda(X)) := \min_{t \in [0,1]} \Big[t + \frac{1}{(1-\beta)m} \sum_{y \in Y} (z_y - t)_+ \Big], \tag{5}$$

where $(u)_+ = \max\{u, 0\}$. Equivalently, sort $z_{(1)} \geq \cdots \geq z_{(m)}$, and let $q = (1-\beta)m$ and $r = \lceil q \rceil$. Then

$$\mathcal{S}_\beta(Y, C_\lambda(X)) = \frac{\sum_{j=1}^{r-1} z_{(j)} + (q - r + 1)\, z_{(r)}}{q}. \tag{6}$$

The level β sets how much of the tail is averaged. At $\beta = 0$ the loss is the average severity $\mathcal{S}_{\text{avg}}$ in (4), and larger β concentrates it on fewer, more severe mismatches, up to the worst label-level severity $\max_{y \in Y} z_y$ once $\beta \geq 1 - 1/m$. The same β thus acts as worst-case severity for patients with few medications and as a tail average for patients with many.

Proposition 3.1 (Properties of the severity loss)**.** *For every $\beta \in [0, 1)$, the loss $\mathcal{S}_\beta(Y, C_\lambda(X))$*

(i) lies in $[0, 1]$;

(ii) is non-increasing in λ;

(iii) equals 0 at $\lambda = 1$.

The same holds for $\mathcal{S}_{\text{avg}} = \mathcal{S}_0$.

These are the conditions of conformal risk control [Angelopoulos et al., 2024]: a bounded loss that is non-increasing in λ and vanishes at the largest threshold.

Remark 3.2 (Connection to FNR control). For $Y \neq \emptyset$,

$$\mathcal{S}_{\text{avg}}(Y, C_\lambda(X)) = \frac{1}{|Y|} \sum_{y \in Y \setminus C_\lambda(X)} s\big(y, C_\lambda(X)\big) \;\leq\; \frac{|Y \setminus C_\lambda(X)|}{|Y|} = \text{FNR}(Y, C_\lambda(X)).$$

CRC-FNR calibrates the loss under the hierarchy-blind cost $s(y, k) = \mathbb{1}\{k \neq y\}$. At a common α on the same calibration data, CSC-Avg therefore selects $\hat{\lambda}_{\text{CSC}} \leq \hat{\lambda}_{\text{FNR}}$, giving nested sets $C_{\hat{\lambda}_{\text{CSC}}}(X) \subseteq C_{\hat{\lambda}_{\text{FNR}}}(X)$. The gap $\text{FNR} - \mathcal{S}_{\text{avg}} = \frac{1}{|Y|} \sum_{y \in Y \setminus C_\lambda(X)} \big(1 - s(y, C_\lambda(X))\big)$ is the credit for near misses. CSC thus meets the same tolerance with smaller sets, and the saving comes from missed medications that have an ATC relative in the set.

3.4 Conformal Severity Control

We now introduce CSC, a post-hoc calibration procedure that selects the threshold λ to control patient-level severity. Given a new patient, our goal is to construct prediction sets whose expected patient-level severity is controlled below a user-specified tolerance α. To achieve this without retraining the predictor or assuming a parametric distribution, we work in the CRC framework under exchangeability [Bates et al., 2021, Angelopoulos et al., 2024]. Using the held-out calibration split, we select a threshold $\hat{\lambda}$ with finite-sample severity control. For a tolerance $\alpha \in (0, 1)$, a level $\beta \in [0, 1)$, and the loss $\mathcal{S}_\beta$ from Section 3.3, the threshold is

$$\hat{\lambda} := \min \left\{ \lambda \in \Lambda : \frac{1}{n+1} \left(\sum_{i=1}^{n} \mathcal{S}_\beta(Y_i, C_\lambda(X_i)) + 1 \right) \leq \alpha \right\}, \quad \Lambda = \{j/400 : j = 0, \ldots, 400\}, \tag{7}$$

with $\min \emptyset := 1$. The $+1$ term accounts for the unseen test patient. Every loss is zero at $\lambda = 1$ by Proposition 3.1(iii), so the feasible set is empty only when $\alpha < 1/(n+1)$, and the fallback then returns the full label set. The prediction set for a new patient is $C_{\hat{\lambda}}(X)$.

Theorem 3.3 (Expected severity control)**.** *Let f be a fixed score function, and let $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ be exchangeable. For any $\beta \in [0, 1)$ and $\alpha \in (0, 1)$, the threshold $\hat{\lambda}$ in* (7) *satisfies*

$$\mathbb{E}\left[\mathcal{S}_\beta(Y, C_{\hat{\lambda}}(X))\right] \leq \alpha, \tag{8}$$

where the expectation is over the calibration data and the new patient.

Theorem 3.3 is the CRC guarantee [Angelopoulos et al., 2024, Theorem 1] applied to $\mathcal{S}_\beta$, with Proposition 3.1 supplying its conditions. The guarantee is marginal: it bounds severity averaged over calibration draws and future patients. For $\beta > 0$, it controls the expected within-patient tail, rather than an upper quantile of severity across the patient population. Exchangeability is supported by the random patient-level split in Section 4 and does not extend to temporal or cross-site shift.

3.5 Algorithmic Summary

Algorithm 1 summarizes CSC. Phase 1 runs once on the calibration split. It scans Λ in ascending order, computes $\mathcal{S}_\beta$ for every calibration patient, and stops at the first λ that meets the tolerance, returning $\hat{\lambda} = 1$ if none does. Phase 2 applies $\hat{\lambda}$ to new patients. Theorem 3.3 holds for any finite grid containing $\lambda = 1$, and refining the grid can only lower $\hat{\lambda}$; we use a step of $1/400$.

Algorithm 1 Conformal Severity Control: Offline Calibration and Online Prediction

Require: Calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$, scoring model f trained without $\mathcal{D}_{\text{cal}}$, tolerance $\alpha \in (0, 1)$, level $\beta \in [0, 1)$, ATC tree of depth D, grid $\Lambda = \{j/400 : j = 0, \ldots, 400\}$.

Ensure: Calibrated threshold $\hat{\lambda}$ and prediction set $C_{\hat{\lambda}}(X)$ for a new patient X.

% Phase 1: Offline Calibration
$\hat{\lambda} \leftarrow 1$ (fallback: full label set)
for $\lambda \in \Lambda$ in ascending order **do**
  for each $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$ **do**
    $C_\lambda(X_i) \leftarrow \{k \in \mathcal{K} : f_k(X_i) \geq 1 - \lambda\}$
    if $Y_i = \emptyset$ **then**
      $\mathcal{S}_i(\lambda) \leftarrow 0$
    else
      $z_{i,y} \leftarrow s(y, C_\lambda(X_i))$ for each $y \in Y_i$, using $s(y, k) = 1 - \text{depth}(\text{LCA}(y, k))/D$
      Sort $z_{i,(1)} \geq \cdots \geq z_{i,(m_i)}$, where $m_i = |Y_i|$; set $q_i = (1 - \beta)m_i$, $r_i = \lceil q_i \rceil$
      $\mathcal{S}_i(\lambda) \leftarrow \left(\sum_{j=1}^{r_i - 1} z_{i,(j)} + (q_i - r_i + 1)\, z_{i,(r_i)}\right)/q_i$
    end if
  end for
  if $\frac{1}{n+1}\left(\sum_{i=1}^{n} \mathcal{S}_i(\lambda) + 1\right) \leq \alpha$ **then**
    $\hat{\lambda} \leftarrow \lambda$; **break**
  end if
end for
% Phase 2: Online Inference for a New Patient
Compute scores $f_k(X)$ for all $k \in \mathcal{K}$
Return $C_{\hat{\lambda}}(X) = \{k \in \mathcal{K} : f_k(X) \geq 1 - \hat{\lambda}\}$

4 Experiment Design

Dataset and Pre-processing. We evaluate on MIMIC-IV, a publicly available de-identified EHR dataset of patients admitted to the emergency department or intensive care unit (ICU) at Beth Israel Deaconess Medical Center in Boston, MA [Johnson et al., 2023]. It contains inpatient EHR records, including diagnoses, procedures, medication prescriptions, and admission/discharge timestamps. Diagnoses and procedures are represented using ICD codes, medications are mapped from NDC to ATC level-4 codes. Dataset statistics and preprocessing are detailed in Appendix C.

Medication Recommendation Problem. We construct patient sequences by treating each inpatient admission as one visit. For patient i, let $X_i = [X_i^{(1)}, \ldots, X_i^{(N_i)}]$ denote the temporally ordered visit sequence, where each visit contains diagnoses, procedures, and medications. We select the last visit $t = N_i$ of each patient as the target visit. The model input includes patient demographics, diagnoses and procedures from visits 1 through t, and medication history from visits 1 through $t - 1$. The prediction target is the medication set prescribed at visit t, denoted by $Y_i^{(t)} \subseteq \mathcal{K}$.

LLM Drug Recommender. We use an LLM-based drug recommendation predictor built on the Llama-based *Meerkat-8B* model [Kim et al., 2025]. The Meerkat-8B LLM takes patient demographics and longitudinal visit history as input using the prompt template provided in Appendix C.4. We obtain sequence embeddings using LLM2Vec [BehnamGhader et al., 2024] and place a multi-label prediction head on top of these embeddings. We fine-tune the model with LoRA adapters on the attention projection layers while keeping the remaining parameters frozen [Hu et al., 2022]. The model is trained with binary cross-entropy (BCE) over medication labels and predicts a score for each medication label. For evaluation, we report precision, recall, and F1.

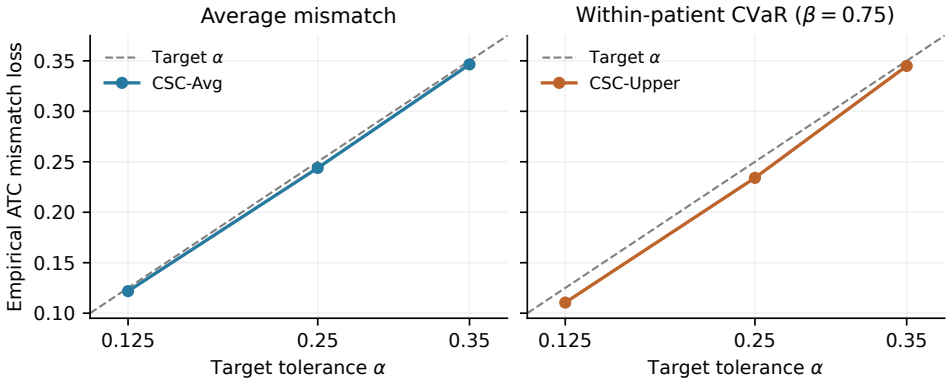


Figure 4: **Empirical ATC taxonomy-based mismatch severity across target tolerances.** For each α, CSC selects $\hat{\lambda}$ on the full calibration split, and the resulting prediction sets are evaluated once on the original test split. The left panel shows average severity (CSC-Avg); the right panel shows within-patient empirical CVaR at $\beta = 0.75$ (CSC-Upper). The dashed line marks $y = \alpha$. Each point is a single-split empirical estimate.

Data Splits and Calibration. We split patients into training, calibration, and test sets in a 60:20:20 ratio. We use patient-level splits, with all visits from the same patient kept in one split, so exchangeability is assumed between calibration and test patient-level examples. The training split is used to fine-tune the LLM-based predictor for drug recommendation. The calibration split is used to select the CSC threshold $\hat{\lambda}$, and the test split is held out for final evaluation.

Calibration Setup. We evaluate CSC under both average severity and upper-quantile severity control at tolerance levels $\alpha \in \{0.125, 0.25, 0.35\}$. Since medications are evaluated at ATC level 4, label-level severity takes values in $\{0, 0.25, 0.5, 0.75, 1\}$. Thus, α specifies the target average patient-level severity. For example, $\alpha = 0.25$ means that the calibrated prediction sets allow an average severity comparable to a one-level ATC mismatch. For upper-quantile severity control, we evaluate $\beta \in \{0.50, 0.75, 0.90\}$, where larger β focuses on fewer high-severity mismatches per patient. The calibrated threshold $\hat{\lambda}$ is selected on the calibration split and then fixed for evaluation on the test split.

Baselines. We evaluate two CSC variants, **CSC-Avg** and **CSC-Upper**, against both conformal and non-conformal baselines. For the conformal baseline, **CRC-FNR** calibrates the multi-label false-negative rate (FNR) rather than the hierarchy-aware mismatch severity. For non-conformal baselines, **Score-threshold** selects all labels whose predicted probability exceeds a global cutoff τ, while **Score-top-k** selects the k highest-scoring labels for each patient. We tune τ or k on the calibration split to match the recall or prediction set size of CSC. **Prevalence-top-k** ignores patient-specific scores and recommends the k most frequent medication labels in the calibration split.

Evaluation Metrics. We report empirical average severity, upper-quantile severity, prediction set size, precision, recall, and F1 on the held-out test split. Severity metrics evaluate the calibrated prediction sets under the ATC hierarchy, while precision, recall, and F1 measure exact label overlap.

5 Results

5.1 Severity Calibration Across User-Specified Tolerances

We evaluate whether CSC achieves empirical severity close to the user-specified tolerance α. For each α, we select $\hat{\lambda}$ on the full calibration split and evaluate the resulting prediction sets once on the original test split. Figure 4 shows that CSC-Avg attains average severity 0.1218, 0.2439, and 0.3465 at $\alpha = 0.125$, 0.25, and 0.35, and CSC-Upper with $\beta = 0.75$ attains within-patient empirical CVaR 0.1105, 0.2341, and 0.3449. All six point estimates lie below their targets. For a fixed prediction set, larger β defines a stricter loss because it averages fewer and higher-severity mismatches; Table 7 reports other values of β, each calibrated separately.

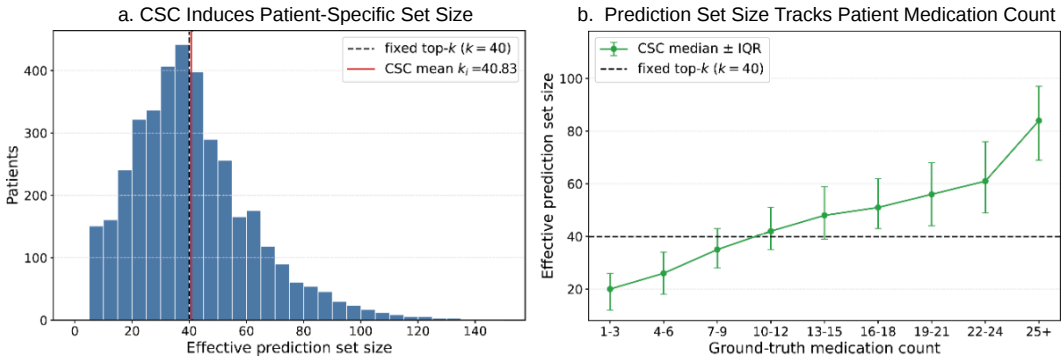


Figure 5: **CSC produces patient-specific prediction sets under a single calibrated threshold $\hat{\lambda}$.** (a) CSC matches fixed Top-k baseline in average set size but produces a broad distribution of k_i across patients. (b) Larger ground-truth medication sets correspond to larger calibrated sets, while fixed Top-k returns the same set size for all patients without considering patient-specific mismatch severity.

Table 1: **Calibration results across target tolerances.** For each α, $\hat{\lambda}$ is selected on the full calibration split, and all other entries are point estimates on the original test split. The controlled loss is average severity for CSC-Avg and within-patient empirical CVaR at $\beta = 0.75$ for CSC-Upper. Precision, recall, and F1 are averaged across patients.

Method	α	$\hat{\lambda}$	Set Size	Controlled Loss	Precision	Recall	F1
CSC-Avg	0.125	0.9325	28.53	0.1218	0.315	0.813	0.438
	0.25	0.8225	16.38	0.2439	0.443	0.661	0.510
	0.35	0.7175	11.66	0.3465	0.522	0.551	0.512
CSC-Upper ($\beta = 0.75$)	0.125	0.9875	62.87	0.1105	0.166	0.945	0.274
	0.25	0.9675	40.83	0.2341	0.240	0.888	0.366
	0.35	0.9425	30.94	0.3449	0.298	0.833	0.423

Table 1 shows the resulting trade-off: for CSC-Avg, increasing α from 0.125 to 0.35 reduces the mean set size from 28.53 to 11.66, raising precision from 0.315 to 0.522 and lowering recall from 0.813 to 0.551. At the same α, CSC-Upper returns larger sets because it controls the worst-covered medications within each patient.

Finite-sample correction. To isolate the $+1$ correction in Eq. (7), we compare CSC-Avg with a plug-in rule that uses the same scores, loss, and grid but selects the smallest λ whose uncorrected calibration loss is at most α. At $\alpha = 0.25$, we draw 100 calibration subsamples without replacement for each $n \in \{25, 50, 100\}$ and evaluate both rules on the same fixed test split. Table 2 shows that the plug-in test loss exceeds α in 38–41% of subsamples, whereas CSC exceeds it in 13–28%. CSC pays for this with larger sets, 2.07 additional labels on average at $n = 25$. With the full calibration split, the two rules nearly coincide. These exceedance fractions are empirical diagnostics on a fixed test split, not failure probabilities of Theorem 3.3, which bounds the expected loss. Appendix D.3 extends the comparison to $n \leq 2{,}000$.

5.2 Patient-Adaptive Set Sizes under a Global Severity Threshold

We next examine the set sizes that a single global threshold induces. For patient i, the threshold $\hat{\lambda}$ yields an effective set size $k_i = |C_{\hat{\lambda}}(X_i)|$ that depends on the patient's score distribution, whereas fixed Top-k assigns the same k to every patient. Figure 5a shows CSC-Upper at $\alpha = 0.25$ and $\beta = 0.75$: its mean set size is close to that of size-matched Top-k ($\bar{k}_i = 40.83$ versus $k = 40$), but k_i varies widely across patients (median 38, interquartile range $[26, 52]$). Figure 5b shows that k_i increases with the ground-truth medication count $|Y_i|$. This variation describes how the calibrated cutoff distributes set sizes across patients; it does not by itself imply lower ATC mismatch at a matched budget, which Table 3 examines.

Table 2: **Effect of the finite-sample correction at $\alpha = 0.25$ (CSC-Avg).** Plug-in omits the $+1$ correction in Eq. (7). For $n \leq 100$, entries are means over 100 calibration subsamples evaluated on the same fixed test split; brackets give empirical 2.5–97.5% quantiles of test loss across subsamples, and Exceed is the fraction of subsamples with test loss above α. The full-calibration row is a single point estimate.

	CSC			Plug-in		
n	Loss [quantiles]	Size	Exceed	Loss [quantiles]	Size	Exceed
25	0.214 [0.154, 0.287]	18.67	13%	0.245 [0.171, 0.319]	16.60	40%
50	0.230 [0.183, 0.284]	17.44	19%	0.245 [0.198, 0.303]	16.45	41%
100	0.238 [0.203, 0.275]	16.80	28%	0.246 [0.210, 0.287]	16.35	38%
3,826	0.244	16.38	–	0.246	16.23	–

Table 3: **Matched baseline diagnostics at $\alpha = 0.25$.** All rules use the same predictor scores. Matching parameters are selected on the calibration split, and all entries are point estimates on the original test split. CRC-FNR matched to CSC's calibration recall or mean set size selects the same cutoff as CSC and returns identical sets. Score-top-k is matched to CSC's mean calibration set size. Bold marks the loss that each CSC variant controls.

Method	Set size	$\widehat{S}_{\text{avg}}$	$\widehat{S}_{0.75}$	Precision	Recall	F1
CSC-Avg (= matched CRC-FNR)	16.38	**0.2439**	0.6247	0.443	0.661	0.510
Score-top-k ($k = 16$)	16.00	0.2206	0.6032	0.414	0.677	0.462
CSC-Upper (= matched CRC-FNR)	40.83	0.0680	**0.2341**	0.240	0.888	0.366
Score-top-k ($k = 40$)	40.00	0.0649	0.2361	0.235	0.889	0.346

5.3 Calibration and Set-Size Trade-offs under Matched Baselines

Table 3 compares CSC with baselines that use the same predictor scores at $\alpha = 0.25$. All matching parameters are selected on the calibration split before test evaluation. CRC-FNR uses the same nested threshold family as CSC; when its target is matched to CSC's calibration recall or mean set size, it selects the same cutoff and returns identical sets. Score-top-k uses a different set-construction rule. At matched mean size, it attains lower average loss than CSC-Avg (0.2206 at $k = 16$ versus 0.2439) and similar within-patient CVaR to CSC-Upper (0.2361 at $k = 40$ versus 0.2341), while CSC attains higher precision and F1 in both comparisons. CSC therefore does not yield lower ATC mismatch at a fixed budget. Its role is to choose the cutoff directly from a prespecified ATC-risk tolerance with the finite-sample correction in Eq. (7); Table 2 isolates the effect of that correction. Appendix D.5 reports all tolerances and the prevalence baseline.

6 Discussion and Limitations

CSC controls an ATC taxonomy-based mismatch cost. The consistency between the mismatch cost and the practical clinical harms need to be further studied. Our evidence covers one single-center dataset, MIMIC-IV, its 300 most frequent ATC-4 labels, and two predictors, an LLM-based one and a non-LLM one (Appendix D.6), and the marginal guarantee assumes exchangeability. Future work includes clinician-grounded evaluation and extending CSC to temporal and cross-hospital distribution shift.

7 Conclusion

We introduced CSC, a post-hoc framework for calibrating medication prediction sets under ATC-based mismatch severity. CSC replaces binary missed-label errors with a hierarchy-aware severity loss and calibrates a single threshold to control expected patient-level severity for any trained scoring model under exchangeability. On MIMIC-IV, CSC converts user-specified severity tolerances into calibrated prediction sets: α sets the allowed severity level, while β controls emphasis on the highest-severity mismatches. The calibrated sets track the target severity levels, adapt their size across patients under one global threshold.

Acknowledgments and Disclosure of Funding

Computing resources were provided via an NVIDIA Academic Grant Program award to Ziyang Song at Ohio University. Zijun Yu acknowledges partial support from an ISM Graduate Scholarship. Chi-Kuang Yeh acknowledges partial support from start-up funds provided by Georgia State University.

References

Zafar Ali, Yi Huang, Irfan Ullah, Junlan Feng, Chao Deng, Nimbeshaho Thierry, Asad Khan, Asim Ullah Jan, Xiaoli Shen, Wu Rui, and Guilin Qi. Deep learning for medication recommendation: A systematic survey. *Data Intell.*, 5(2):303–354, May 2023.

Hang Lv, Zixuan Guo, Yanchao Tan, Wanzi Shao, Hengyu Zhang, and Carl Yang. Awaken the giant: Activating llms via deep model guidance for boundary-aware medication recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '26, page 1030–1041, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722585. doi: 10.1145/3770854.3780297. URL `https://doi.org/10.1145/3770854.3780297`.

Xinyu Wang, Hanwei Wu, Zhenghan Tai, Sicheng Lyu, Qincheng Lu, Ziyu Zhao, Jijun Chi, Jingrui Tian, Xiao-Wen Chang, and Ziyang Song. Saferx-agent: A knowledge-grounded multi-agent framework for safe and explainable medication recommendation, 2026. URL `https://arxiv.org/abs/2605.29146`.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit. Med.*, 4(1):4, January 2021.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature*, 01 2019. doi: 10.1038/s42256-018-0004-1.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and J. Sun. Gamenet: Graph augmented memory networks for recommending medication combination. volume 33, pages 1126–1133, 07 2019. doi: 10.1609/aaai.v33i01.33011126.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and J. Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 3735–3741, 08 2021. doi: 10.24963/ijcai.2021/514.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, March 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68:1–34, 12 2021. doi: 10.1145/3478535.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=33XGfHLtZg`.

Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. Leave no patient behind: Enhancing medication recommendation for rare disease patients. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 533–542, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657785.

WHO Collaborating Centre for Drug Statistics Methodology WHOCC. *Guidelines for ATC classification and DDD assignment*. Oslo, 2026.

Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17:449–476, 02 2007. doi: 10.1111/j.1467-9965.2007.00311.x.

Jiayi Huang, Amirmohammad Farzaneh, and Osvaldo Simeone. Optimized certainty equivalent risk-controlling prediction sets, 2026. URL https://arxiv.org/abs/2602.13660.

Ziyang Song, Qincheng Lu, Hao Xu, He Zhu, David Buckeridge, and Yue Li. Timelygpt: Extrapolatable transformer pre-training for long-term time-series forecasting in healthcare. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400713026. doi: 10.1145/3698587.3701364. URL https://doi.org/10.1145/3698587.3701364.

Ziyang Song, Qincheng Lu, He Zhu, David Buckeridge, and Yue Li. Trajgpt: Irregular time-series representation learning of health trajectory. *IEEE Journal of Biomedical and Health Informatics*, 30(5):3888–3899, 2026a. doi: 10.1109/JBHI.2025.3620205.

Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S. Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3970–3980, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539089. URL https://doi.org/10.1145/3534678.3539089.

Zijun Yao, Bin Liu, Fei Wang, Daby Sow, and Ying Li. Ontology-aware prescription recommendation in treatment pathways using multi-evidence healthcare data. *ACM Trans. Inf. Syst.*, 41(4), April 2023.

Mengxuan Sun, Jinghao Niu, Xuebing Yang, Yifan Gu, and Wensheng Zhang. Cehmr: Curriculum learning enhanced hierarchical multi-label classification for medication recommendation. *Artif. Intell. Med.*, 143(C), September 2023. ISSN 0933-3657. doi: 10.1016/j.artmed.2023.102613. URL https://doi.org/10.1016/j.artmed.2023.102613.

Weicong Tan, Weiqing Wang, Xin Zhou, Wray Buntine, Gordon Bingham, and Hongzhi Yin. OntoMedRec: Logically-pretrained model-agnostic ontology encoders for medication recommendation. *World Wide Web*, 27(3), May 2024.

Janette Vazquez and Julio C Facelli. Conformal prediction in clinical medical sciences. *J. Healthc. Inform. Res.*, 6(3):241–252, September 2022.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Dr.Regina Barzilay. Conformal prediction sets with limited false positives. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6514–6532. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/fisch22a.html.

Yong-Cui Wang, Shi-Long Chen, Nai-Yang Deng, and Yong Wang. Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, 29(10):1317–1324, 2013. doi: 10.1093/bioinformatics/btt158.

Justin Bohn, James P. Gilbert, Christopher Knoll, David M. Kern, and Patrick B. Ryan. Large-scale empirical identification of candidate comparators for pharmacoepidemiological studies. *Drug Safety*, 48(11):1229–1241, 2025. doi: 10.1007/s40264-025-01569-y.

R Rockafellar and Stan Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 01 2000.

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. doi: 10.1038/s41597-022-01899-x.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ Digit. Med.*, 8(1):240, May 2025.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ziyang Song, Lia Shen, Yixuan Li, Qincheng Lu, Weihao Li, Haoran Xu, He Zhu, Xinyu Wang, and Yue Li. Smi: Ontology-grounded semantic ids for hierarchy-aware medical concept representation, 2026b. URL https://arxiv.org/abs/2610.18672.

Yixuan Li, Weihao Li, and Ziyang Song. Hcoe: Hyperbolic clinical ontology embeddings from biomedical language models, 2026. URL https://arxiv.org/abs/2609.30763.

Yabin Kuang and Minzhu Xie. DrugDoctor: enhancing drug recommendation in cold-start scenario via visit-level representation learning and training. *Briefings in Bioinformatics*, 25(6):bbae464, 2024. doi: 10.1093/bib/bbae464.

A Notations of Variables

Table 4: Summary of notation used throughout the paper.

Notation	Description
Spaces and label sets	
$\mathcal{X}$	Input space.
$\mathcal{K}$	Medication label space.
$\mathcal{Y}$	Output space of label sets, e.g., subsets of $\mathcal{K}$.
$k \in \mathcal{K}$	Candidate or predicted medication label.
$Y \subseteq \mathcal{K}$	Ground-truth medication label set for a patient.
$y \in Y$	Single ground-truth medication label.
Random variables, observations, and calibration data	
$X \in \mathcal{X}$	Patient/model input, viewed as a random variable.
$\{(X_i, Y_i)\}_{i=1}^{n}$	Exchangeable calibration dataset, viewed as random input–label pairs.
x_i	Observed features of visit i in a patient's longitudinal record.
n	Number of calibration examples.
i	Calibration-example index.
$\mathbb{E}[\cdot]$	Expectation with respect to future exchangeable data.
Prediction sets and thresholding	
$f_k(X)$	Model score assigned to label k for input X.
$\lambda \in [0, 1]$	Threshold parameter controlling the prediction set size.
$C_\lambda(X)$	Prediction set at threshold λ.
λ_1, λ_2	Thresholds used to express the nestedness of prediction sets.
$\hat{\lambda}$	Calibrated threshold for the generic or upper-quantile severity loss.
$\hat{\lambda}_{\text{avg}}$	Calibrated threshold for average-severity control.
Hierarchy-aware severity	
$s : \mathcal{K} \times \mathcal{K} \to [0, 1]$	Hierarchy-aware pairwise severity function.
$s(y, k)$	Severity of predicting label k when the ground-truth label is y.
$\text{LCA}(y, k)$	Lowest common ancestor of y and k in the ATC hierarchy.
$\text{depth}(\cdot)$	Depth of a node in the ATC hierarchy.
$s(y, C_\lambda(X))$	Severity of the closest label in $C_\lambda(X)$ to the true label y.
Patient-level losses and conformal calibration	
$\ell(C_\lambda(X), Y)$	Generic bounded monotone loss used in the conformal calibration framework.
$\mathcal{L}$	Generic patient-level severity loss used by CSC.
$\mathcal{L}_i(\lambda)$	Patient-level loss of calibration example i at threshold λ.
$\beta \in [0, 1)$	Severity quantile level
$\mathcal{S}_\beta(Y, C_\lambda(X))$	Patient-level upper-tail severity loss at severity quantile level β.
$\mathcal{S}_{\text{avg}}(Y, C_\lambda(X))$	Patient-level average severity loss. This is a special case of $\mathcal{S}_\beta(Y, C_\lambda(X))$ where $\beta = 0$.
$\alpha \in (0, 1)$	User-specified target tolerance for the population loss or severity.
Drug recommendation task notation	
$[X_i^{(1)}, \ldots, X_i^{(N_i)}]$	Temporally ordered patient sequence.
$Y_i^{(t)}$	Target medications set at visit t.

B Proofs of the Propositions and Main Theorem

B.1 Proof of Proposition 3.1

Proof of Proposition 3.1. Part (i). For nonempty Y, define

$$F_\beta(t,\lambda) := t + \frac{1}{(1-\beta)|Y|}\sum_{y\in Y}\big(s(y,C_\lambda(X)) - t\big)_+.$$

Since $F_\beta(t,\lambda) \geq 0$ for $t \in [0,1]$ and $s(y,C_\lambda(X)) \leq 1$,

$$0 \leq \mathcal{S}_\beta(Y,C_\lambda(X)) = \min_{t\in[0,1]} F_\beta(t,\lambda) \leq F_\beta(1,\lambda) = 1.$$

For empty Y, the loss is zero by convention.

Part (ii). For fixed $Y \neq \emptyset$ and β, each label-level severity is non-increasing in λ. Thus, for $\lambda_1 \leq \lambda_2$,

$$F_\beta(t,\lambda_2) \leq F_\beta(t,\lambda_1) \quad \text{for every } t \in [0,1].$$

Taking minima over the same interval gives

$$\begin{aligned}\mathcal{S}_\beta(Y,C_{\lambda_2}(X)) &= \min_{t\in[0,1]} F_\beta(t,\lambda_2) \\ &\leq \min_{t\in[0,1]} F_\beta(t,\lambda_1) = \mathcal{S}_\beta(Y,C_{\lambda_1}(X)).\end{aligned}$$

For empty Y, the loss remains zero.

Part (iii) holds because $C_1(X) = \{k : f_k(X) \geq 0\} = \mathcal{K} \supseteq Y$, so $s(y,C_1(X)) = s(y,y) = 0$ for every $y \in Y$, and the loss of an all-zero collection is 0. □

B.2 Proof of Theorem 3.3

Proof of Theorem 3.3. We establish the result by verifying that $\mathcal{S}_\beta(Y,C_\lambda(X))$ satisfies the conditions of conformal risk control [Angelopoulos et al., 2024], then applying that framework carefully to handle the selection of $\hat{\lambda}$.

Step 1: Loss properties. By Proposition 3.1, the severity satisfies two properties required by the conformal risk control framework.

1. **Boundedness.** $\mathcal{S}_\beta(y,C_\lambda(x)) \in [0,1]$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$ and all $\lambda \in \Lambda$.
2. **Monotonicity.** $\mathcal{S}_\beta(y,C_{\lambda_1}(x)) \geq \mathcal{S}_\beta(y,C_{\lambda_2}(x))$ whenever $\lambda_1 \leq \lambda_2$, since larger λ induces larger prediction sets, $C_{\lambda_1}(x) \subseteq C_{\lambda_2}(x)$, and severity is non-increasing in set size.

Step 2: Calibration on a finite grid. The threshold $\hat{\lambda}$ in (7) is selected from the finite grid $\Lambda = \{j/400 : j = 0,\ldots,400\}$, which contains $\lambda = 1$, with the fallback $\hat{\lambda} = 1$ when no $\lambda \in \Lambda$ satisfies the calibration condition. The argument below uses only that Λ is finite and contains 1. The threshold $\hat{\lambda}$ is a function of the calibration data $\{(X_i,Y_i)\}_{i=1}^n$ only.

Step 3: Expected loss on the test point. Define the augmented empirical risk

$$\widehat{R}(\lambda;\mathcal{D}_{n+1}) := \frac{1}{n+1}\sum_{i=1}^{n+1}\mathcal{S}_\beta(Y_i,C_\lambda(X_i)),$$

where

$$\mathcal{D}_{n+1} = \{(X_i,Y_i)\}_{i=1}^{n+1}, \qquad (X_{n+1},Y_{n+1}) = (X,Y).$$

Consider the auxiliary threshold

$$\lambda^* = \min\Big\{\lambda \in \Lambda : \widehat{R}(\lambda;\mathcal{D}_{n+1}) \leq \alpha\Big\}.$$

By Proposition 3.1(iii), $\widehat{R}(1;\mathcal{D}_{n+1}) = 0 \leq \alpha$, so this set is nonempty and λ^* is well defined.

Since λ^* is a symmetric function of the augmented dataset $\mathcal{D}_{n+1}$, and the samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are jointly exchangeable, the random variables

$$\mathcal{S}_\beta(Y_i, C_{\lambda^*}(X_i)), \qquad i = 1, \ldots, n+1,$$

are exchangeable. Therefore,

$$\begin{aligned}\mathbb{E}[\mathcal{S}_\beta(Y, C_{\lambda^*}(X))] &= \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}[\mathcal{S}_\beta(Y_i, C_{\lambda^*}(X_i))] \\ &= \mathbb{E}\Big[\widehat{R}(\lambda^*; \mathcal{D}_{n+1})\Big] \le \alpha. \end{aligned} \tag{9}$$

It remains to compare λ^* with the calibrated threshold $\hat{\lambda}$ defined in (7). By Step 1,

$$0 \le \mathcal{S}_\beta(Y, C_\lambda(X)) \le 1 \qquad \text{for all } \lambda.$$

Hence, for every $\lambda \in \Lambda$,

$$\widehat{R}(\lambda; \mathcal{D}_{n+1}) = \frac{1}{n+1}\sum_{i=1}^{n+1}\mathcal{S}_\beta(Y_i, C_\lambda(X_i)) \le \frac{1}{n+1}\left(\sum_{i=1}^{n}\mathcal{S}_\beta(Y_i, C_\lambda(X_i)) + 1\right).$$

Therefore, every λ satisfying the calibration condition in (7) is also feasible for the definition of λ^*, which implies

$$\lambda^* \le \hat{\lambda}.$$

If no $\lambda \in \Lambda$ satisfies the calibration condition, then $\hat{\lambda} = 1 \ge \lambda^*$, so the inequality holds in both cases.

Finally, by monotonicity of $\mathcal{S}_\beta(Y, C_\lambda(X))$ with respect to λ,

$$\mathcal{S}_\beta(Y, C_{\hat{\lambda}}(X)) \le \mathcal{S}_\beta(Y, C_{\lambda^*}(X)) \qquad \text{almost surely.}$$

Taking expectations and applying (9) gives

$$\mathbb{E}\big[\mathcal{S}_\beta(Y, C_{\hat{\lambda}}(X))\big] \le \mathbb{E}[\mathcal{S}_\beta(Y, C_{\lambda^*}(X))] \le \alpha,$$

which proves (8).

Step 4: The $\beta = 0$ case. When $\beta = 0$, the severity functional reduces to the average severity, $\mathcal{S}_0 = \mathcal{S}_{\text{avg}}$. The boundedness and monotonicity properties in Step 1 hold for $\mathcal{S}_{\text{avg}}$ by Proposition 3.1, so Steps 2–3 apply verbatim, yielding (8) for all $\beta \in [0, 1)$. □

C Dataset Details

C.1 ATC Hierarchy and Severity Interpretation

The World Health Organization's Anatomical Therapeutic Chemical (ATC) system organizes medications into a five-level hierarchy, from broad anatomical categories to specific chemical substances [WHOCC, 2026]. In our setting, this hierarchy contains 14 anatomical main groups, 90 therapeutic subgroups, 248 pharmacological subgroups, 840 chemical subgroups, and 5,495 chemical substances. Each ATC code can therefore be viewed as a path in a rooted tree, where longer shared prefixes indicate greater therapeutic similarity [Song et al., 2026b, Li et al., 2026]. For example, `A10BA02` denotes metformin, and its successive prefixes `A`, `A10`, `A10B`, and `A10BA` encode its ancestor categories in the ATC hierarchy. In our experiments, medications are represented at **ATC level-4**, as described in Appendix C.3.

This ATC hierarchy gives a direct interpretation of the severity function in Eq. (2). For ATC-4 labels, a lowest common ancestor at depth 4 (same ATC-4 code) yields severity 0; depth 3 (same pharmacological subgroup) yields severity $1/4$; depth 2 (same therapeutic class) yields severity $1/2$; depth 1 (same anatomical group) yields severity $3/4$; and depth 0 (different top-level groups) yields severity 1. Thus, prediction sets are penalized less for including medications closer to the ground-truth in the ATC hierarchy and more for including only distant medications.

Table 5: Structure of the ATC medication hierarchy used in this study. *# Concepts* denotes the number of unique medication concepts at each level.

Level	Meaning	# Concepts	Example Code	Example Code Description
1	Anatomical main group	14	`A`	Alimentary tract and metabolism
2	Therapeutic main group	90	`A10`	Drugs used in diabetes
3	Pharmacological subgroup	248	`A10B`	Blood glucose lowering drugs, excluding insulins
4	Chemical subgroup	840	`A10BA`	Biguanides
5	Chemical substance	5,495	`A10BA02`	Metformin

C.2 Dataset and Cohort Statistics

We used a large-scale, de-identified MIMIC-IV EHR dataset from Beth Israel Deaconess Medical Center (BIDMC), released via PhysioNet under a credentialed data use agreement [Johnson et al., 2023]. This dataset contains longitudinal inpatient records, including diagnoses, procedures, medication prescriptions, and admission/discharge timestamps. Patient identifiers are removed and all timestamps are shifted by a patient-specific offset. This preserves relative temporal ordering within each patient sequence while preventing alignment of absolute calendar time across patients. We construct visit-based patient sequences by treating each inpatient admission as a visit and ordering visits chronologically by admission time. Table 6 reports basic cohort statistics after preprocessing.

Table 6: Basic statistics of the pre-processed MIMIC-IV EHR dataset. "#" denotes counts and "/patient" denotes the average number of records per-patient.

Dataset	#patients	#visits	#visits/patient	#conditions/patient	#procedures/patient	#drugs/patient
MIMIC-IV	19,135	35,974	1.88	21.74	4.70	43.89

C.3 Preprocessing and Concept Normalization

We applied the same preprocessing pipeline to the MIMIC-IV dataset, including selection criteria, code normalization, and data splitting, yielding consistent visit-based patient EHR sequences. We applied the same normalization procedure across cohorts to ensure a fair comparison.

Cohort construction and splitting. We treat each inpatient admission as a visit and represent each patient as a chronologically ordered sequence of visits. We perform patient-level training/calibration/test splits so that all visits from the same patient appear in only one split, avoiding information leakage across splits.

Diagnosis and procedure codes. For each visit, we extract diagnoses and procedures from the corresponding MIMIC-IV tables and represent them as unordered sets of codes. We retain the original ICD coding systems (ICD-9-CM and ICD-10-CM for diagnoses; ICD-9 and ICD-10-PCS for procedures) and preserve the recorded ICD version to avoid conflating codes across systems. We map ICD-9/10 diagnosis and procedure codes to CCS categories, and de-duplicate repeated codes within each visit. Specifically, 11,736 ICD-9-CM and 72,446 ICD-10-CM diagnosis codes are mapped to 285 CCS condition categories, while 4,670 ICD-9-PROC and 79,758 ICD-10-PCS procedure codes are mapped to 231 CCS procedure categories, yielding 285 condition and 231 procedure concepts in total.

Medication mapping and normalization. We follow the ATC-L4 medication normalization setting used in SafeRx-Agent [Wang et al., 2026]. Medications are extracted from the prescription tables, where drugs are recorded as National Drug Codes (NDCs). We standardize them using a two-stage mapping pipeline: **NDC** $\rightarrow$ **RxNorm (RxCUI)** $\rightarrow$ **ATC**. We retain **ATC level-4** codes to preserve meaningful therapeutic and chemical structure. Unmappable NDC entries (e.g., non-drug supplies, institution-specific packaging codes, or missing references) are excluded from the medication representation. Finally, we retain the 300 most frequent ATC level-4 medication codes as the prediction vocabulary. The severity loss is evaluated over this retained medication vocabulary.

Visit-level feature construction. Each visit x_i is represented by three unordered sets: diagnosis codes, procedure codes, and mapped ATC-4 medication codes. These per-visit sets are then composed into a time-ordered patient sequence $\{x_1, \ldots, x_t\}$ by admission time. Basic cohort statistics after preprocessing are reported in Table 6.

C.4 LLM Prompt Template

Prompt Template for Drug Recommendation

Demographics: Insurance: <INSURANCE>; Language: <LANGUAGE>;
Religion: <RELIGION>; Marital status: <MARITAL_STATUS>; Ethnicity: <ETHNICITY>.
The patient has <VISIT_NUM> ICU visits.
In visit 1, diagnoses: <DIAG_NAME>, ...; procedures: <PROC_NAME>, ...; medications: <MED_NAME>, ...
...
Current visit: diagnoses: <DIAG_NAME>, ...; procedures: <PROC_NAME>, ...
Task: Predict the medications to prescribe for the current visit. Output a multi-label set of **4-level ATC codes**. Choose any number of codes from the medication vocabulary. Note that a patient may have multiple medications.
Output format: A multi-label set of ATC level-4 codes.

C.5 Computing Resources

Experiments for training the LLM predictor are conducted on an NVIDIA DGX Spark system with 128 GB of memory. Training each LLM takes 2.45 hours per epoch. For MIMIC-IV, we further split the training portion into training and validation subsets. The model is fine-tuned for up to 10 epochs, with early stopping based on validation performance. Mixed-precision training (FP16) is used to improve training efficiency.

D Additional Results

D.1 LLM and CSC Prediction Set Matching Analysis

We provide additional implementation details and result discussion for the prediction set matching analysis shown in Figure 2. The goal of this analysis is to characterize whether treatment-associated medications are exactly matched, or instead matched by therapeutically close or distant medications in the prediction set. Beyond exact-overlap metrics, this analysis measures the hierarchical distance between each ground-truth medication and its closest predicted medication under the ATC taxonomy. To achieve this, we construct four disease cohorts representing distinct clinical categories and identify a set of representative treatment medications for each disease category. We compare the prediction sets generated by the baseline LLM drug recommender with the CSC-calibrated prediction sets constructed from the same LLM predictor. For each ground-truth treatment medication, we identify the closest predicted medication in the ATC hierarchy and categorize the match as exact, close, moderate, or distant according to the shared hierarchy depth.

Cohort Construction. We analyze four disease cohorts: cardiovascular disease, diabetes, respiratory disease, and neurology. The cardiovascular cohort contains 346 patients and corresponds to hypertension or chronic cardiovascular disease. The diabetes cohort contains 11 patients and corresponds to type 2 diabetes or metabolic syndrome. The respiratory cohort contains 29 patients and corresponds to COPD or asthma. The neurology cohort contains 25 patients and corresponds to epilepsy or chronic neurologic disease. These cohorts are analyzed independently, so the same patient may appear in more than one cohort because comorbidities are common.

For each cohort, we evaluate only ground-truth medications that belong to ATC branches associated with the target disease category, not all medications in the patient EHR data. For cardiovascular disease, we use C02, C03, C07, C08, and C09. For diabetes, we use A10. For respiratory disease,

we use R03. For neurology, we use N03A. Across the four cohorts, each patient has an average of 2.91 cardiovascular-associated, 1.55 diabetes-associated, 1.07 respiratory-associated, and 1.36 neurology-associated ground-truth medication labels, respectively.

Prediction Sets from the LLM Drug Recommender and CSC. We use the LLM-based drug predictor described in Section 4 as the base prediction model and apply the CSC framework to its medication prediction scores. This yields two prediction sets for comparison. The first is the LLM prediction set obtained by thresholding the LLM-predicted medication probabilities at 0.25. The second is the CSC-calibrated prediction set obtained from the same LLM prediction scores using CSC-Avg with target severity tolerance $\alpha = 0.125$. For each selected target medication, we compare it with all medications in the prediction set and select the one with the closest ATC hierarchy match. We then categorize the best match as exact, close, moderate, or distant according to its ATC hierarchy-aware mismatch severity: exact hits have severity 0, close non-exact matches have severity in $(0, 0.25]$, moderate matches have severity 0.5, and distant mismatches have severity at least 0.75.

Overall Results Across Disease Cohorts. Across all evaluated disease-associated medication labels from the four cohorts, CSC achieves better best-match outcomes than the LLM prediction sets. The exact-hit rate increases from 51.56% for the LLM prediction sets to 94.13% for the CSC-calibrated prediction sets. The exact-or-close match rate increases from 57.16% to 95.96%. The distant mismatch rate decreases from 34.86% to 0.83%, and the mean best-match severity decreases from 0.328 to 0.0268. These results show that CSC shifts prediction sets away from distant mismatches and toward exact or close medication matches.

Macro-Averaged Results. Because the cardiovascular cohort contains many more labels than the other cohorts, we also compute macro-averaged results across the four disease groups. The macro-average exact-hit rate increases from 29.69% for the LLM prediction sets to 90.26% for the CSC-calibrated prediction sets. The macro-average distant mismatch rate decreases from 51.64% to 1.79%, and the mean best-match severity decreases from 0.490 to 0.0394. These results show that the improvement is not driven only by the large cardiovascular cohort; CSC consistently shifts disease-associated medication matches away from distant mismatches and toward exact hits across disease areas.

Cardiovascular Cohort Results. In the cardiovascular cohort, the exact-hit rate increases from 53.77% for the LLM prediction sets to 94.54% for the CSC-calibrated prediction sets. The distant mismatch rate decreases from 33.33% to 0.69%. In absolute counts, the LLM prediction sets produce 542 exact hits, 43 close non-exact matches, 87 moderate matches, and 336 distant mismatches. The CSC-calibrated prediction sets produce 953 exact hits, 15 close non-exact matches, 33 moderate matches, and only 7 distant mismatches. These results show that CSC substantially increases exact hits and nearly removes distant ATC-branch mismatches in the largest cohort.

Diabetes Cohort Results. In the diabetes cohort, the exact-hit rate increases from 11.76% for the LLM prediction sets to 88.24% for the CSC-calibrated prediction sets. The distant mismatch rate decreases from 70.59% to 0%. In absolute counts, the LLM prediction sets produce 2 exact hits, 3 close non-exact matches, and 12 distant mismatches. The CSC-calibrated prediction sets produce 15 exact hits, 2 close non-exact matches, and no distant mismatches. Although this cohort is small, these results show that CSC substantially increases exact hits and removes distant ATC-branch mismatches.

Respiratory Cohort Results. In the respiratory cohort, the exact-hit rate increases from 3.23% for the LLM prediction sets to 87.10% for the CSC-calibrated prediction sets. The distant mismatch rate decreases from 96.77% to 6.45%. In absolute counts, the LLM prediction sets produce 1 exact hit and 30 distant mismatches. The CSC-calibrated prediction sets produce 27 exact hits, 2 moderate matches, and 2 distant mismatches. These results show that CSC produces the largest reduction in distant ATC-branch mismatches in the respiratory cohort.

Neurology Cohort Results. In the neurology cohort, the exact-hit rate increases from 50.00% for the LLM prediction sets to 91.18% for the CSC-calibrated prediction sets. The distant mismatch rate decreases from 5.88% to 0%. In absolute counts, the LLM prediction sets produce 17 exact hits, 15 close non-exact matches, and 2 distant mismatches. The CSC-calibrated prediction sets produce

Table 7: **Effect of β at $\alpha = 0.25$: fixed threshold versus separate calibration.** The fixed-threshold column keeps the CSC-Upper threshold calibrated at $\beta = 0.75$ ($\hat{\lambda} = 0.9675$, mean set size 40.83) and changes only β in the loss. The separate-calibration columns select a new threshold for each β. Losses are patient-averaged within-patient empirical CVaR; all entries are point estimates on the original test split.

	Fixed threshold	Separate calibration		
β	Test loss	$\hat{\lambda}$	Test loss	Set size
0.50	0.1335	0.9275	0.2463	27.48
0.75	0.2341	0.9675	0.2341	40.83
0.90	0.3521	0.9825	0.2364	54.08

31 exact hits, 3 close non-exact matches, and no distant mismatches. These results show that CSC increases exact hits and removes the remaining distant ATC-branch mismatches in this cohort.

Transitions from Distant Mismatches to Exact Hits or Close Matches. We further measure how often the same ground-truth medication changes from a distant mismatch under the LLM prediction sets to an exact hit or close match under the CSC-calibrated prediction sets. This transition occurs for 30.65% of cardiovascular labels, 70.59% of diabetes labels, 83.87% of respiratory labels, and 5.88% of neurology labels. The mean severity reduction is largest in respiratory disease, where it reaches 0.831. The corresponding reductions are 0.544 for diabetes, 0.286 for cardiovascular disease, and 0.140 for neurology. These results show that CSC not only increases exact hits, but also moves many high-severity distant mismatches into exact or low-severity categories.

Key Findings. This prediction-set matching analysis shows why exact-overlap evaluation is not sufficient for medication recommendation. Exact-overlap metrics measure whether a ground-truth medication is exactly included in the prediction set; however, they do not distinguish a therapeutically close medication match in the ATC hierarchy from a distant match. The proposed hierarchy-aware mismatch severity captures this difference by measuring how far the closest predicted medication is from each ground-truth medication. The disease-cohort results show that the LLM prediction sets often leave disease-associated medications matched only to distant ATC branches, especially in the diabetes and respiratory cohorts. In contrast, the CSC-calibrated prediction sets substantially reduce distant mismatches and shift the best-match distribution toward exact hits and close matches. These findings show that CSC improves the structure of prediction errors, not only the number of exact hits.

D.2 Effect of Upper-Quantile Level β

Increasing β has two effects. For a fixed prediction set, it moves the loss toward the most severe mismatches within each patient; under separate calibration, it also changes the selected threshold. Table 7 separates the two at $\alpha = 0.25$. Calibration with $\beta = 0.75$ selects $\hat{\lambda} = 0.9675$. Holding this threshold fixed gives mean set size 40.83 for every β, and the patient-averaged empirical CVaR rises from 0.1335 at $\beta = 0.50$ to 0.2341 at $\beta = 0.75$ and 0.3521 at $\beta = 0.90$. The last value exceeds 0.25 because this threshold was calibrated for $\beta = 0.75$. Calibrating each β separately instead selects larger thresholds as β grows, with mean set sizes 27.48, 40.83, and 54.08 and losses 0.2463, 0.2341, and 0.2364, all below the target. Larger β thus gives stronger protection against the worst-covered medications at the cost of larger sets.

D.3 Finite-Sample Correction across Calibration Sizes

This appendix extends Table 2 to larger calibration sizes. Both rules use the same predictor scores, loss, and grid Λ; CSC applies the $+1$ correction in Eq. (7), and the plug-in rule omits it. For each n, we draw 100 calibration subsamples without replacement from the 3,826 calibration patients. Within a repetition, the subsamples for different n are nested prefixes of one random permutation, and both rules use the same subsample. All rules are evaluated on the same fixed test split. As n grows, the correction shrinks and the two rules converge. For $n \geq 1{,}000$, subsamples drawn from the same 3,826 patients overlap substantially, so their spread understates the variability of independent calibration sets.

Table 8: **Finite-sample correction across calibration sizes at $\alpha = 0.25$ (CSC-Avg).** For $n \leq 2{,}000$, loss and size are means over 100 calibration subsamples on a fixed test split; brackets give empirical 2.5–97.5% quantiles of test loss, and Exceed is the fraction of subsamples with test loss above α. Exceed is a fixed-test empirical diagnostic, not a failure probability of Theorem 3.3. The row $n = 3{,}826$ uses the full calibration split once.

	CSC			Plug-in		
n	Loss [quantiles]	Size	Exceed	Loss [quantiles]	Size	Exceed
25	0.214 [0.154, 0.287]	18.67	13%	0.245 [0.171, 0.319]	16.60	40%
50	0.230 [0.183, 0.284]	17.44	19%	0.245 [0.198, 0.303]	16.45	41%
100	0.238 [0.203, 0.275]	16.80	28%	0.246 [0.210, 0.287]	16.35	38%
250	0.242 [0.220, 0.262]	16.53	23%	0.245 [0.224, 0.266]	16.36	30%
500	0.244 [0.229, 0.259]	16.36	20%	0.246 [0.229, 0.261]	16.28	26%
1,000	0.245 [0.236, 0.254]	16.33	13%	0.246 [0.236, 0.254]	16.29	18%
2,000	0.245 [0.239, 0.249]	16.31	2%	0.245 [0.240, 0.249]	16.29	2%
3,826	0.244	16.38	–	0.246	16.23	–

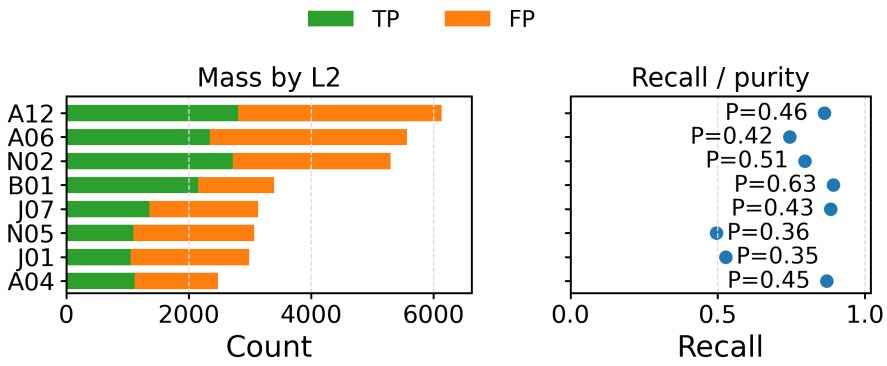


Figure 6: ATC level-2 branch composition for average-severity CSC at $\alpha = 0.25$. The left panel shows true positives and false positives among the most frequently selected ATC branches, while the right panel reports branch-level recall and precision.

D.4 ATC Hierarchy Branch Statistics

We further inspect how CSC allocates prediction mass across the ATC hierarchy. For each ATC level-2 branch, we aggregate medication-level true positives and false positives among predicted medications. We also report branch-level recall and purity, where purity is the branch-level precision $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$. These plots are diagnostic only.

Figure 6 shows the branch-level composition for CSC-Avg at $\alpha = 0.25$. Most prediction mass is concentrated in common ATC level-2 branches, including A12 mineral supplements, A06 drugs for constipation, N02 analgesics, B01 antithrombotic agents, and J07 vaccines. CSC-Avg keeps prediction sets relatively compact, so branch-level purity remains moderate in several common branches, but recall varies across medication families.

Figure 7 shows the corresponding branch-level composition for CSC-Upper with $\alpha = 0.25$ and $\beta = 0.75$. Because CSC-Upper controls the worst-covered medications within each patient, it selects larger prediction sets and expands coverage across more ATC branches. This increases branch-level recall in many branches, but also introduces more false positives and lowers purity. The pattern is consistent with the intended trade-off of upper-tail severity control: reducing high-severity misses requires more conservative prediction sets.

D.5 Matched Baseline Comparisons

Table 9 extends Table 3 to all three tolerances. All matching parameters are selected on the calibration split and applied once to the original test split, so the test set sizes of matched baselines can differ slightly from those of CSC. Size matching selects the k whose mean calibration set size is closest

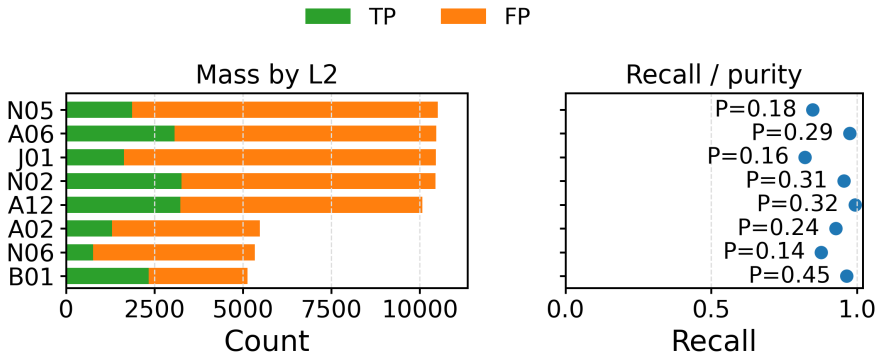


Figure 7: ATC level-2 branch composition for upper-quantile severity CSC at $\alpha = 0.25$ and $\beta = 0.75$. Compared with average-severity calibration, upper-tail calibration selects larger sets and increases branch-level recall in many medication families, at the cost of lower branch-level purity.

to that of CSC, and recall matching does the same for patient-averaged recall; ties go to the more inclusive candidate. Score-top-k ranks medications by predictor score for each patient, with ties broken by label index. Prevalence-top-k returns the k medications most frequent in the calibration split to every patient.

CRC-FNR is not listed separately: in all six settings, matching its target to CSC's calibration recall or mean set size selects CSC's cutoff and returns identical sets. Score-threshold uses the same threshold family, and at CSC's cutoff it reproduces CSC's sets. Score-top-k attains lower average loss than CSC-Avg at every tolerance under both matching criteria. Its within-patient CVaR differs from that of CSC-Upper by at most 0.012, lower at one setting and higher at the others. CSC attains higher patient-averaged precision and F1 than Score-top-k in every comparison. These diagnostics therefore show no fixed-budget advantage for CSC in ATC mismatch; they report the set-size and exact-overlap cost of the operating point that CSC selects for a prespecified tolerance. Prevalence-top-k, which ignores patient-specific scores, has much higher loss at matched size and needs much larger sets to match recall.

D.6 Application to DrugDoctor-Based Scores

We apply CSC unchanged to scores from a non-LLM recommender based on DrugDoctor [Kuang and Xie, 2024]. The scores were supplied as precomputed matrices for the same calibration and test patients. We verified that patient IDs and label matrices match those of the LLM predictor, and the authors confirmed the ordering of the 300-label vocabulary; only the scores differ. The loss matrix, grid, and calibration rule are unchanged. Table 10 shows that all six test losses lie below their targets, and CSC-Upper again selects larger sets with higher recall and lower precision. The supplied files do not document whether DrugDoctor's training and model selection excluded the calibration and test patients, so we report these results as an empirical application and do not invoke Theorem 3.3 for them.

D.7 Sensitivity to the Taxonomy Cost

We recalibrate CSC-Avg at $\alpha = 0.25$ under three monotone transformations of the normalized LCA distance $d = s(y, k)$ in Eq. (2), namely $\sqrt{d}$, d, and d^2, which change the penalty for nonzero mismatches but preserve its ordering. We also use a branch-dependent cost built from the ATC tree. For a node v, let $N(v)$ be its number of ATC-5 descendants and $I(v) = -\log\{N(v)/N(\text{root})\}$; the cost between labels u and v is $I(u) + I(v) - 2I(\text{LCA}(u, v))$, normalized by its maximum over the 300-label vocabulary. Among the 302 label pairs whose LCA has depth 3, the equal-edge cost is 0.25 for every pair, whereas this cost ranges from 0.081 to 0.449. Every cost is bounded and zero for exact matches, so the loss stays monotone and the calibration procedure is unchanged. For label `L01XC`, which is absent from the hierarchy file, we keep its code-prefix path and count one pseudo-leaf.

Table 9: **Matched baseline diagnostics at all tolerances.** Matching parameters are selected on the calibration split, and all entries are point estimates on the original test split. CRC-FNR matched to CSC's calibration recall or mean set size returns the same sets as CSC in every block. "Size" and "recall" mark the matching criterion. Median [IQR] summarizes set sizes across test patients; fixed-k rules give every patient the same size. Bold marks the loss that CSC controls in each block.

Rule	Mean size	Median [IQR]	$\widehat{S}_{\text{avg}}$	$\widehat{S}_{0.75}$	Precision	Recall	F1
CSC-Avg, $\alpha = 0.125$ ($\hat{\lambda} = 0.9325$)							
CSC-Avg (= matched CRC-FNR)	28.53	27 [17, 37]	**0.1218**	0.3833	0.315	0.813	0.438
Score-top-k (size, $k = 28$)	28.00	–	0.1108	0.3754	0.303	0.821	0.405
Score-top-k (recall, $k = 27$)	27.00	–	0.1171	0.3917	0.310	0.813	0.410
Prevalence-top-k (size, $k = 28$)	28.00	–	0.2146	0.6128	0.246	0.655	0.327
Prevalence-top-k (recall, $k = 50$)	50.00	–	0.1043	0.3529	0.175	0.817	0.270
CSC-Avg, $\alpha = 0.25$ ($\hat{\lambda} = 0.8225$)							
CSC-Avg (= matched CRC-FNR)	16.38	14 [8, 23]	**0.2439**	0.6247	0.443	0.661	0.510
Score-top-k (size, $k = 16$)	16.00	–	0.2206	0.6032	0.414	0.677	0.462
Score-top-k (recall, $k = 15$)	15.00	–	0.2360	0.6262	0.426	0.658	0.464
Prevalence-top-k (size, $k = 16$)	16.00	–	0.3700	0.7778	0.314	0.504	0.347
Prevalence-top-k (recall, $k = 28$)	28.00	–	0.2146	0.6128	0.246	0.655	0.327
CSC-Avg, $\alpha = 0.35$ ($\hat{\lambda} = 0.7175$)							
CSC-Avg (= matched CRC-FNR)	11.66	9 [5, 17]	**0.3465**	0.7573	0.522	0.551	0.512
Score-top-k (size, $k = 12$)	12.00	–	0.2913	0.6965	0.467	0.593	0.466
Score-top-k (recall, $k = 10$)	10.00	–	0.3388	0.7445	0.501	0.541	0.463
Prevalence-top-k (size, $k = 12$)	12.00	–	0.4455	0.8341	0.341	0.426	0.339
Prevalence-top-k (recall, $k = 19$)	19.00	–	0.3098	0.7164	0.296	0.555	0.348
CSC-Upper ($\beta = 0.75$), $\alpha = 0.125$ ($\hat{\lambda} = 0.9875$)							
CSC-Upper (= matched CRC-FNR)	62.87	58 [41, 80]	0.0307	**0.1105**	0.166	0.945	0.274
Score-top-k (size, recall; $k = 62$)	62.00	–	0.0289	0.1121	0.164	0.945	0.264
Prevalence-top-k (size, $k = 62$)	62.00	–	0.0744	0.2642	0.149	0.865	0.241
Prevalence-top-k (recall, $k = 105$)	105.00	–	0.0246	0.0912	0.098	0.944	0.171
CSC-Upper ($\beta = 0.75$), $\alpha = 0.25$ ($\hat{\lambda} = 0.9675$)							
CSC-Upper (= matched CRC-FNR)	40.83	38 [26, 52]	0.0680	**0.2341**	0.240	0.888	0.366
Score-top-k (size, recall; $k = 40$)	40.00	–	0.0649	0.2361	0.235	0.889	0.346
Prevalence-top-k (size, $k = 40$)	40.00	–	0.1488	0.4776	0.202	0.758	0.296
Prevalence-top-k (recall, $k = 70$)	70.00	–	0.0622	0.2243	0.136	0.887	0.224
CSC-Upper ($\beta = 0.75$), $\alpha = 0.35$ ($\hat{\lambda} = 0.9425$)							
CSC-Upper (= matched CRC-FNR)	30.94	29 [19, 40]	0.1073	**0.3449**	0.298	0.833	0.423
Score-top-k (size, $k = 31$)	31.00	–	0.0960	0.3330	0.283	0.843	0.390
Score-top-k (recall, $k = 30$)	30.00	–	0.1010	0.3475	0.289	0.836	0.395
Prevalence-top-k (size, $k = 31$)	31.00	–	0.1946	0.5794	0.233	0.690	0.320
Prevalence-top-k (recall, $k = 55$)	55.00	–	0.0966	0.3302	0.163	0.834	0.257

Table 11 shows that every cost stays below its own target, while the cutoff, set size, and recall change with the cost. Losses on different scales are compared only with the target in their own row. These costs remain taxonomy-based and do not measure clinical harm.

Table 10: **CSC with DrugDoctor-based scores.** Each row calibrates on all 3,826 calibration patients and reports point estimates on the 3,827 test patients. Bold marks the loss that each CSC variant controls.

Method	α	$\hat{\lambda}$	$\widehat{S}_{\text{avg}}$	$\widehat{S}_{0.75}$	Set size	Precision	Recall	F1
CSC-Avg	0.125	0.9375	**0.1197**	0.3708	43.86	0.204	0.816	0.316
CSC-Avg	0.25	0.8575	**0.2464**	0.6091	24.73	0.309	0.659	0.399
CSC-Avg	0.35	0.7700	**0.3462**	0.7316	16.73	0.385	0.541	0.419
CSC-Upper	0.125	0.9825	0.0312	**0.1103**	93.81	0.105	0.940	0.185
CSC-Upper	0.25	0.9600	0.0740	**0.2436**	58.86	0.159	0.876	0.262
CSC-Upper	0.35	0.9450	0.1054	**0.3337**	47.63	0.191	0.834	0.301

Table 11: **Taxonomy-cost sensitivity for CSC-Avg at $\alpha = 0.25$.** Each cost is calibrated separately on the full calibration split, and entries are point estimates on the original test split. Each test loss is on its own scale and is compared only with the 0.25 target.

Cost	$\hat{\lambda}$	Test loss	Set size	Precision	Recall	F1
$\sqrt{d}$	0.8550	0.2464	18.64	0.412	0.700	0.499
d (equal-edge)	0.8225	0.2439	16.38	0.443	0.661	0.510
d^2	0.7700	0.2440	13.68	0.485	0.603	0.515
Branch-dependent IC	0.6950	0.2477	10.94	0.537	0.529	0.508

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: a post-hoc, model-agnostic CSC framework, an hierarchy-based mismatch severity based on ATC taxonomy, calibration of prediction sets under average severity and upper-quantile severity. These claims are supported by the formal severity definitions, the conformal calibration procedure, the expected severity control theorem, and the empirical calibration and baseline comparisons.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its current scope, noting that it focuses on medication recommendation and frames extensions to other hierarchical clinical tasks, such as ICD-based diagnosis, and to non-healthcare domains, such as finance, as future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical results state the required assumptions explicitly. Proposition 3.1 establishes boundedness, monotonicity, and zero loss at the full label set which are the properties needed for CSC. Theorem 3.3 states the exchangeability assumption on the calibration and test examples and specifies the valid ranges of α and β. Complete proofs are provided in the appendix, including the exchangeability argument used to obtain marginal expected severity control.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the full experimental pipeline needed to reproduce the main results, including the MIMIC-IV preprocessing steps, the mapping from NDC codes to ATC level-4 labels, the patient-level 60:20:20 train/calibration/test split, and the evaluation metrics. It also describes the base LLM drug recommender, including the Meerkat-8B backbone, the LLM2Vec embedding setup, the LoRA fine-tuning strategy, and the BCE training objective, and the prompt template. For CSC, the paper specifies the calibration protocol, the severity-control settings over α and β, and how the baselines are tuned on the calibration split for fair comparison. The appendix further reports implementation details such as cohort construction, computational resources, and the matching-analysis setup used in the additional experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper uses public but credentialed-access clinical dataset MIMIC-IV, whose redistribution is restricted by data-use agreements. We provide implementation details in the appendix, but the current submission does not include an anonymized code release.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C provides the experimental details needed to interpret and reproduce the results, including patient-level cohort construction and the 60:20:20 train/calibration/test split. It also documents the preprocessing pipeline, including diagnosis and procedure code handling, the NDC-to-RxNorm-to-ATC medication mapping, and the ATC level-4 codes. The paper further describes the LLM-based predictor, including the Meerkat-8B backbone, LLM2Vec embeddings, BCE training objective, LoRA fine-tuning strategy, prompt template, calibration protocol, and evaluation metrics. The appendix also reports additional implementation details, including the CSC severity-control settings, baseline

tuning on the calibration split, computing resources, training epochs, early stopping, and mixed-precision training. Together, these details make the training and evaluation setup sufficiently transparent.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports uncertainty estimates for the main severity calibration experiment. Figure 4 and Table 1 report single-split point estimates. Appendix D.3 reports empirical 2.5–97.5% quantiles of the test loss across 100 calibration subsamples drawn without replacement ($n_{\text{cal}} \in \{25, \ldots, 2000\}$) on a fixed test split. These quantiles describe calibration-sampling variability; they are not confidence intervals for the expected loss.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C.5 summarizes the compute resources used for the main experiments, including the hardware used for fine-tuning the LLM predictor, the memory available, the approximate training time per epoch, the maximum number of epochs, early stopping, and mixed-precision training. The paper also notes that CSC is an offline calibration procedure applied after the base predictor is trained, so its computational cost is negligible relative to fine-tuning and inference for the LLM-based recommender.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses de-identified, credentialed-access clinical datasets and does not involve intervention on patients, crowdsourcing, or collection of new human-subject data. We found no deviation from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive societal impact through its intended use in medication recommendation, where CSC can make prediction sets more clinically useful by reducing high-severity therapeutic mismatches and allowing users to tune the tolerated severity level. The paper also acknowledges the current scope of the work, noting that the experiments focus on medication recommendation and that extensions to other hierarchical clinical tasks and non-healthcare domains are future work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a deployable model, a pretrained or fine-tuned LLM checkpoint. Therefore, the work does not introduce a new released asset that would require controlled access or additional misuse-mitigation safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the original sources of the external assets used in our study, including the MIMIC-IV dataset, the ATC medication hierarchy, and the pretrained models and components used in the base LLM recommender. We use these assets under their respective access conditions, licenses, and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission does not currently distribute a new dataset, model checkpoint, or standalone software package alongside the paper. While no newly released asset is provided with the submission, we plan to release code after publication,

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, participant recruitment, or direct interaction with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study uses de-identified, controlled-access EHR datasets and does not collect new human-subject data or interact with participants. Dataset access is governed by the original data-use agreements.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: The paper describes its use of LLMs in the experimental pipeline. Specifically, the base drug recommender is built on the Meerkat-8B LLM, with LLM2Vec embeddings, a multi-label prediction head, LoRA fine-tuning, and a prompt template provided in Appendix C.4. The paper makes clear that the LLM is used as the underlying predictive backbone, while the core methodological contribution is the post-hoc CSC calibration procedure rather than a new LLM architecture or prompting method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.