CoREL: Conditional Relative Evidence Learning for Multi-Institutional Medication Recommendation

Anonymous authors
Paper under double-blind review

Abstract

Medication recommendation in real-world clinical settings must account for heterogeneity in prescribing practices across healthcare institutions. Existing methods typically fit observed prescriptions directly, potentially conflating patient-dependent prescribing evidence with institution-specific prescribing tendencies. We propose CoREL, a multi-institutional medication recommendation framework that separates shared evidence learning from institution-specific prescription calibration. Under an additive decomposition of prescription decisions, we introduce conditional relative evidence learning to eliminate institution–medication tendencies while enabling shared patient-dependent prescribing evidence to be learned across institutions. Dual-level calibration then establishes a global prescribing reference and estimates institution-specific tendencies, recovering institution-conditioned prescription probabilities. An inference-time refinement procedure further balances prescription support against known harmful drug–drug interactions (DDIs). Experiments on eICU, MIMIC-IV, and a real-world two-center inpatient dataset demonstrate improved recommendation accuracy and lower DDI rates, with particularly pronounced gains in multi-institutional settings. Code is available at https://anonymous.4open.science/r/CoREL-01C5.

1 Introduction

The widespread adoption of electronic health records (EHRs) has provided large-scale longitudinal data Adler-Milstein & Jha (2017); Shickel et al. (2017), facilitating the development and application of artificial intelligence (AI) for data-driven clinical decision support Rajkomar et al. (2018); Yu et al. (2018); Sutton et al. (2020). Given the central role of pharmacotherapy in clinical practice Barnett et al. (2016), learning-based medication recommendation has received substantial attention. Early approaches leveraged longitudinal patient histories for personalized recommendation Bhoi et al. (2023); Shang et al. (2019); Wu et al. (2022), while subsequent work incorporated drug molecular structures and drug–drug interaction (DDI) information to improve both recommendation accuracy and safety Yang et al. (2021; 2023); Du et al. (2026). Together, these advances have led to consistent improvements in predictive performance on standard EHR benchmarks, highlighting the potential of medication recommendation to support real-world prescribing decisions.

To translate this potential into greater clinical utility, however, medical AI must contend with heterogeneity in clinical practice across healthcare institutions Yang et al. (2022); Lasko et al. (2024). For medication recommendation in particular, a patient's clinical state does not uniquely determine the final prescription. Multiple clinically viable regimens may exist for the same patient, while the eventual medication choices also depend on local drug availability, prescribing habits, and other institution-specific factors beyond patient characteristics Pedersen et al. (2017); Schachtner et al. (2002); Ozaki et al. (2022). Figure 1(a) provides a concrete example from eICU benchmarks Pollard et al. (2018): omeprazole and esomeprazole, both proton-pump inhibitors with overlapping clinical use in acid-related disease Richter et al. (2001); Katz et al. (2022), exhibit markedly different prescribing patterns across institutions. Consequently, prescribing differences across institutions reflect not only differences in patients' clinical states but also how medication choices are made within different local practices. Accounting for this institutional heterogeneity is therefore impor-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

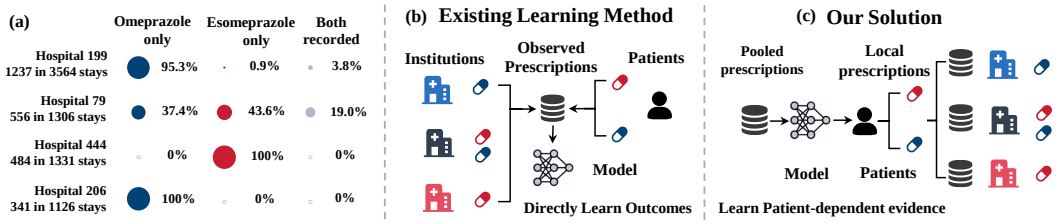

Figure 1: (a) Marked institutional variation in the use of omeprazole and esomeprazole across eICU institutions. (b–c) Existing methods may entangle patient-dependent evidence with institution-specific prescribing practice in multi-institutional setting, while CoREL explicitly separates the two.

tant for developing medication recommendation models that better reflect real-world practice and provide context-appropriate recommendations across diverse clinical settings.

Existing medication recommendation methods do not fully address this institutional heterogeneity. As shown in Figure 1(b), although these methods differ in patient representation and their use of drug knowledge, they typically use observed prescriptions as prediction targets, directly learning a mapping from patient clinical states to the medications ultimately prescribed. In multi-institutional settings, however, observed prescriptions reflect the combined influence of patient-dependent prescribing evidence and institution-specific prescribing tendencies. Without explicitly distinguishing these contributions, a model may incorporate local prescribing tendencies into the predictive mapping shared across institutions, potentially compromising recommendation performance within individual institutions. More accurate reproduction of observed prescriptions alone therefore does not ensure that a model can effectively account for institutional heterogeneity.

A more appropriate formulation of multi-institutional medication recommendation explicitly distinguishes the roles of patient-dependent evidence and institutional tendencies during learning. Our premise is that patient-dependent prescribing evidence is broadly shared across institutions because it is grounded in general pharmacological and pathophysiological relationships between patient conditions and medications. Cross-institutional sharing should therefore focus on this evidence, rather than complete prescribing patterns that also encode local practice. Institutional tendencies, in contrast, should neither be indiscriminately absorbed into shared patient–medication relationships nor be discarded at prediction time. Instead, they should be estimated separately for each institution and explicitly combined with shared patient-dependent evidence to form prescriptions that account for both the patient's clinical state and local prescribing practice.

Building on this formulation, we propose CoREL, a multi-institutional medication recommendation framework that combines shared patient-dependent evidence learning with institution-specific prescription calibration. Specifically, we introduce conditional relative evidence learning, which uses within-institution patient comparisons to separate patient-dependent evidence from local prescribing baselines, enabling shared evidence learning across institutions. To recover institution-conditioned prescription probabilities from this evidence, we introduce dual-level calibration that establishes a global prescribing reference for each medication and then estimates institution-specific tendencies. This global-to-local calibration incorporates local prescribing variation while preserving the learned evidence differences between patients for each medication. Finally, an inference-time DDI-aware refinement procedure is designed to balance prescription support against interaction risk, discouraging the co-selection of medications with known harmful interactions.

Our contributions are summarized as follows:

- We introduce a structured formulation of multi-institutional medication recommendation that distinguishes shared patient-dependent evidence from institution-specific prescribing tendencies, linking cross-institutional knowledge sharing with locally calibrated prescription prediction.
- We develop a conditional relative learning objective that enables patient-dependent prescribing evidence to be learned separately from institutional tendencies, together with dual-level calibration that recovers institution-conditioned prescription probabilities.
- Experiments on two public benchmarks, eICU and MIMIC-IV, and an additional real-world two-center inpatient dataset demonstrate that CoREL improves recommendation accuracy and reduces DDI rates, with particularly pronounced gains in multi-institutional settings.

2 Related Work

Medication Recommendation. Existing EHR-based medication recommendation research has largely focused on single-center settings, with advances primarily driven by patient representation learning and drug knowledge integration. On the patient side, early work models individual clinical states for personalized recommendation Zhang et al. (2017), while subsequent approaches incorporate longitudinal clinical and prescription histories to better characterize patient conditions Shang et al. (2019); Wu et al. (2022); Bhoi et al. (2023). On the drug side, molecular and substructure-level information is incorporated to enrich medication representations Yang et al. (2021; 2023); Du et al. (2026); Feng & Du (2026). Complementing these directions, set-to-set modeling treats patient conditions and prescriptions as unordered sets Tan et al. (2022). Despite these advances, their predictive objectives remain centered on fitting observed prescription labels, with limited explicit consideration of institution-specific prescribing tendencies. This limitation becomes more salient when jointly learning from institutions with heterogeneous prescribing practices. In this paper, CoREL adopts a structured formulation that distinguishes patient-dependent evidence from institutional tendencies to combine cross-institutional knowledge sharing with locally calibrated recommendation.

Learning from Relative Comparisons. Relative comparisons provide relational supervision for learning representations and predictive functions. Relative metric learning methods learn distance functions or embeddings from similarity comparisons Schultz & Joachims (2003); Schroff et al. (2015), while pairwise ranking methods learn scoring functions from relevance or preference relations in retrieval and recommendation Burges et al. (2005); Rendle et al. (2012). When observations share group-level baseline effects, the construction of comparisons becomes important for distinguishing within-group relationships from between-group variation. Conditional likelihood methods provide a statistical basis for this distinction by estimating covariate-dependent relationships while conditioning out group-specific intercepts Cox (1958); Neuhaus & McCulloch (2006). CoREL builds on this conditional perspective to construct medication-specific patient comparisons within each institution, learning shared prescribing evidence separately from institutional prescribing baselines. This evidence is then combined with locally estimated tendencies through global-to-local calibration to obtain institution-conditioned prescription probabilities.

3 Problem Formulation

Electronic Health Records. Let $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|}\}$, $\mathcal{P} = \{p_1, \ldots, p_{|\mathcal{P}|}\}$, and $\mathcal{M} = \{m_1, \ldots, m_{|\mathcal{M}|}\}$ denote the diagnosis, procedure, and medication vocabularies, respectively. A patient q has a sequence of visits $V_q^{(t)} = (D_q^{(t)}, P_q^{(t)}, M_q^{(t)})$, where $D_q^{(t)} \subseteq \mathcal{D}$, $P_q^{(t)} \subseteq \mathcal{P}$, and $M_q^{(t)} \subseteq \mathcal{M}$ denote the diagnosis, procedure, and medication codes recorded at visit t. We further denote by $\mathbf{c}_q$ the patient-level demographic and physiological attributes.

DDI Matrix. Drug–drug interactions are represented by a symmetric matrix $A \in \{0,1\}^{|\mathcal{M}|\times|\mathcal{M}|}$, where $A_{i,j} = 1$ indicates a known harmful interaction between medications m_i and m_j.

Multi-institutional Medication Recommendation. At the prescription decision for visit t, the observed clinical context of patient q is $x_q^{(t)} = \left(\mathbf{c}_q, \{D_q^{(\tau)}, P_q^{(\tau)}\}_{\tau=1}^{t}, \{M_q^{(\tau)}\}_{\tau=1}^{t-1}\right)$. The target prescription is encoded as $\mathbf{y}_q^{(t)} \in \{0,1\}^{|\mathcal{M}|}$, with $y_{q,m}^{(t)} = 1$ iff $m \in M_q^{(t)}$. Each target prescription is observed at a healthcare institution $e_q^{(t)} \in \mathcal{E} = \{e_1, \ldots, e_K\}$. The goal is to learn a shared medication recommendation model across institutions that predicts $\widehat{\mathbf{y}}_q^{(t)} = F\left(x_q^{(t)}, e_q^{(t)}\right) \in \{0,1\}^{|\mathcal{M}|}$.

4 Method

Figure 2 presents the overall framework of CoREL. Our method follows a shared-evidence learning and local-calibration paradigm: it first learns medication-specific patient-dependent prescribing evidence shared across institutions, and then combines this evidence with institution-specific prescribing tendencies to recover locally calibrated prescription predictions. The following sections describe the decomposition, evidence learning, and calibration and inference procedures in detail.

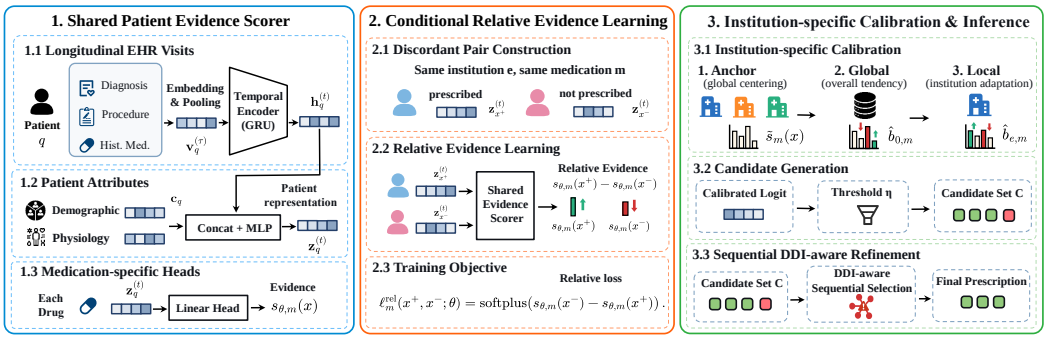

Figure 2: Overview of CoREL. (1) A shared scorer produces medication-specific patient evidence; (2) conditional relative learning estimates this evidence from within-institution patient comparisons; and (3) institution-specific calibration recovers local prescription probabilities, followed by DDI-aware refinement for final recommendation.

4.1 Decomposing Patient Evidence and Institutional Tendency

We first decompose prescribing decisions into two components: patient-dependent prescribing evidence and institution-specific prescribing tendency. The former captures how a patient's observed clinical context supports the use of a medication, and is assumed to be largely shared across institutions because it is grounded in general pathophysiological and pharmacological relationships between patient conditions and medications. The latter captures systematic institution-level variation in the use of individual medications, which reflects how local prescribing practice shifts the baseline propensity to prescribe a medication. Accordingly, we model prescribing decisions as a medication-specific institutional shift on top of a shared patient-dependent evidence.

Formally, let X, E, and Y_m denote the clinical-context, institution, and prescription-indicator random variables, respectively. For medication m, we model its marginal prescription probability as

$$p^\star_{e,m}(x) := \Pr(Y_m = 1 \mid X = x, E = e) = \sigma\big(f^\star_m(x) + b^\star_{e,m}\big)\,, \tag{1}$$

where $\sigma(\cdot)$ is the logistic sigmoid, $f^\star_m(x)$ represents the patient-dependent prescribing evidence shared across institutions, $b^\star_{e,m}$ captures the institution-specific prescribing tendency.

4.2 Conditional Relative Evidence Learning

The decomposition in Eq. (1) does not by itself make its two components separately observable as each observed prescription label reflects their combined effect. Directly fitting absolute prescription probabilities would still entangle patient-dependent prescribing evidence with institution-specific prescribing tendencies. To isolate the patient-dependent component, we introduce *Conditional Relative Evidence Learning (CREL)*, which transforms absolute prescription supervision into relative evidence through conditional comparison.

From Absolute to Relative Prescribing Evidence. The institutional tendency is shared across patients within the same institution for a given medication. This allows us to isolate patient-dependent variation by comparing prescription odds across patient contexts. Specifically, for two patient contexts x and x' observed at the same institution e, Eq. (1) implies

$$\frac{p^\star_{e,m}(x)/(1-p^\star_{e,m}(x))}{p^\star_{e,m}(x')/(1-p^\star_{e,m}(x'))} = \exp(f^\star_m(x) - f^\star_m(x'))\,, \tag{2}$$

where the common institutional term $b^\star_{e,m}$ cancels. Therefore, although the absolute patient-dependent evidence is entangled with the institutional tendency, its difference across patient contexts can be isolated. We refer to $f^\star_m(x) - f^\star_m(x')$ as *relative prescribing evidence*.

Conditional Relative Learning. The prescription probability $p^\star_{e,m}(x)$ in Eq. (2) is latent, whereas the observed data provide only binary prescription outcomes. CREL therefore converts relative prescribing evidence into a learnable signal by conditioning on discordant patient pairs, where exactly one of two patients from the same institution receives medication m.

Specifically, for the same pair of patient contexts (x, x'), let Y_m and Y'_m denote their prescription indicators. Conditioning on $Y_m + Y'_m = 1$, the probability that medication m is prescribed to the patient with context x is

$$\Pr(Y_m = 1 \mid Y_m + Y'_m = 1, X = x, X' = x', E = e) = \sigma(f^\star_m(x) - f^\star_m(x')) \,. \tag{3}$$

This conditioning serves two purposes. First, the common institution-specific tendency is eliminated from the conditional probability. Second, it expresses the relative prescribing evidence in terms of an observable pairwise outcome—which patient in the discordant pair receives medication m. The resulting conditional probability therefore provides a likelihood for the observed pairwise prescription outcome, enabling direct learning of patient-dependent relative prescribing evidence. The derivation of Eq. (3) is provided in Appendix A.

Parameterizing Shared Prescribing Evidence. Eq. (3) establishes a learnable relation for the relative prescribing evidence. We further parameterize the underlying patient-dependent function $f^\star_m(x)$ with a single prescribing evidence model shared across all institutions.

For patient q, diagnosis, procedure, and historical medication codes at each visit are embedded and pooled into a visit representation, and the resulting sequence is encoded longitudinally as

$$\mathbf{v}_q^{(\tau)} = \left[\phi_{\mathcal{D}}(D_q^{(\tau)}); \phi_{\mathcal{P}}(P_q^{(\tau)}); \phi_{\mathcal{M}}(\overline{M}_q^{(\tau)})\right], \qquad \mathbf{h}_q^{(t)} = \mathrm{GRU}_\theta\left(\mathbf{v}_q^{(1)}, \ldots, \mathbf{v}_q^{(t)}\right), \tag{4}$$

where $\overline{M}_q^{(\tau)} = M_q^{(\tau)}$ for $\tau < t$ and $\overline{M}_q^{(t)} = \varnothing$. The longitudinal representation is then combined with patient-level attributes as $\mathbf{z}_q^{(t)} = \mathrm{MLP}_\theta([\mathbf{h}_q^{(t)}; \phi_{\mathcal{C}}(\mathbf{c}_q)])$. The medication-specific prescribing score is then defined as $s_{\theta,m}(x_q^{(t)}) = \mathbf{w}_m^\top \mathbf{z}_q^{(t)}$, which serves as the model estimate of $f^\star_m(x_q^{(t)})$. The patient context encoder and medication-specific scoring functions are shared across institutions, with no institution-specific parameter introduced into this component.

Substituting $s_{\theta,m}$ for the $f^\star_m$ in Eq. (3) yields the final CREL training objective. For a discordant pair, let x^+ and x^- denote the contexts of the prescribed and unprescribed patients, respectively. The negative conditional log-likelihood is

$$\ell^{\mathrm{rel}}_m(x^+, x^-; \theta) = -\log \sigma\big(s_{\theta,m}(x^+) - s_{\theta,m}(x^-)\big) = \mathrm{softplus}\big(s_{\theta,m}(x^-) - s_{\theta,m}(x^+)\big) \,. \tag{5}$$

Balanced Multi-institutional Objective. The number of discordant pairs can vary substantially across institutions and medications. Let $\mathcal{P}_{e,m}$ denote the set of admissible discordant patient pairs for medication m at institution e, and let $\mathcal{M}_e^+ = \{m : |\mathcal{P}_{e,m}| > 0\}$. We train the shared evidence model using the institution- and medication-balanced objective

$$\mathcal{L}_{\mathrm{rel}}(\theta) = \frac{1}{|\mathcal{E}^+|} \sum_{e \in \mathcal{E}^+} \frac{1}{|\mathcal{M}_e^+|} \sum_{m \in \mathcal{M}_e^+} \frac{1}{|\mathcal{P}_{e,m}|} \sum_{(x^+, x^-) \in \mathcal{P}_{e,m}} \ell^{\mathrm{rel}}_m(x^+, x^-; \theta), \tag{6}$$

where $\mathcal{E}^+ = \{e : |\mathcal{M}_e^+| > 0\}$. This hierarchical averaging prevents institutions or medications with more available pairs from dominating the shared evidence model.

4.3 Institution-Specific Prescription Calibration

CREL learns the relative prescribing evidence of different patients for the same medication, but does not determine its absolute level. To recover institution-conditioned prescription probabilities, we perform prescription calibration in two steps. We first anchor the learned patient-dependent evidence to a common reference level over the pooled population, and then estimate medication-specific prescribing tendencies globally and for each institution.

Anchoring Patient-dependent Prescribing Evidence. Because the CREL objective depends only on score differences, adding an arbitrary medication-specific constant to all patient scores leaves the objective unchanged. We therefore center the learned prescribing evidence for each medication over the pooled observed population. Let $\mathcal{S}_e$ denote the training observed prescription records from institution e and $\mathcal{S} = \bigcup_{e \in \mathcal{E}} \mathcal{S}_e$. For medication m, we compute

$$\mu_m = \frac{1}{|\mathcal{S}|} \sum_{(x, \mathbf{y}) \in \mathcal{S}} s_{\hat{\theta},m}(x), \qquad \widetilde{s}_m(x) = s_{\hat{\theta},m}(x) - \mu_m. \tag{7}$$

Thus, the patient-dependent evidence for each medication has zero mean over the pooled population. This transformation preserves all relative evidence learned by CREL while establishing a common reference level for calibrating absolute prescribing tendencies.

Institution-specific Local Calibration. Given the anchored patient-dependent evidence, we first estimate the global prescribing tendency of medication m using records from all institutions:

$$\hat{b}_{0,m} = \arg\min_{b} \frac{1}{|\mathcal{S}|} \sum_{(x,\mathbf{y})\in\mathcal{S}} \mathrm{BCE}\left(y_m, \sigma(\widetilde{s}_m(x) + b)\right). \tag{8}$$

The global tendency $\hat{b}_{0,m}$ calibrates the shared patient-dependent evidence to the overall prescribing level of medication m across institutions. We then calibrate the institution-specific prescribing tendency $b_{e,m}$ from the local prescription records of institution e:

$$\hat{b}_{e,m} = \arg\min_{b} \left\{ \sum_{(x,\mathbf{y})\in\mathcal{S}_e} \mathrm{BCE}\left(y_m, \sigma(\widetilde{s}_m(x) + b)\right) + \frac{\lambda_b}{2}\left(b - \hat{b}_{0,m}\right)^2 \right\}. \tag{9}$$

The shrinkage term $\frac{\lambda_b}{2}(b - \hat{b}_{0,m})^2$ encourages the institution-specific tendency to remain close to the global tendency when local observations are limited, while allowing sufficient local evidence to shift the prescribing level toward the practice of institution e. Here, $\lambda_b > 0$ is a hyperparameter controlling the strength of shrinkage. Importantly, calibration modifies only the medication-specific prescribing tendency and leaves the patient-dependent prescribing evidence unchanged.

4.4 Training and Inference

Training. The model is trained in two stages. In the first stage, we construct medication-specific discordant pairs within each institution and jointly optimize the shared prescribing evidence model using Eq. (6). After convergence, the shared evidence model is fixed, and the prescribing scores over all training contexts are centered according to Eq. (7). In the second stage, prescribing tendencies are calibrated sequentially. We first estimate the global tendency $\hat{b}_{0,m}$ for each medication from the pooled training records, and then estimate the institution-specific tendency $\hat{b}_{e,m}$ from the records of institution e with shrinkage toward $\hat{b}_{0,m}$. Since each tendency is a scalar parameter, both are efficiently solved using Newton's method. For a calibration set $\mathcal{A}$ and shrinkage reference $\bar{b}$, defining $p_i^{(k)} = \sigma(\widetilde{s}_m(x_i) + b^{(k)})$, the update is

$$b^{(k+1)} = b^{(k)} - \frac{\sum_{i\in\mathcal{A}}(p_i^{(k)} - y_{i,m}) + \lambda(b^{(k)} - \bar{b})}{\sum_{i\in\mathcal{A}} p_i^{(k)}(1 - p_i^{(k)}) + \lambda}. \tag{10}$$

We first obtain $\hat{b}_{0,m}$ by applying Eq. (10) to the pooled records $\mathcal{A} = \mathcal{S}$ with $\lambda = 0$. We then obtain $\hat{b}_{e,m}$ using the same update on $\mathcal{A} = \mathcal{S}_e$, with $\bar{b} = \hat{b}_{0,m}$ and $\lambda = \lambda_b$.

Inference With DDI-aware Refinement. For a patient context x from institution e, we first obtain the calibrated prescription probability $\widehat{p}_{e,m}(x) = \sigma\left(\widetilde{s}_m(x) + \widehat{b}_{e,m}\right)$. Let $\ell_m = \widetilde{s}_m(x) + \widehat{b}_{e,m}$ denote the calibrated prescription logit. Given a global prediction threshold η, selected on the validation set and shared across all medications, we define

$$u_m = \ell_m - \mathrm{logit}(\eta), \qquad \mathcal{C}_0 = \{m \in \mathcal{M} : u_m \geq 0\}, \tag{11}$$

where $\mathcal{C}_0$ denotes the medication candidates selected by the calibrated model before DDI refinement.

We then construct the final prescription sequentially. Starting from an empty retained set $\mathcal{R} = \varnothing$, the selection score of each remaining candidate is $r_m(\mathcal{R}) = u_m - \lambda_{\mathrm{ddi}} \sum_{n\in\mathcal{R}} A_{m,n}$, where $\lambda_{\mathrm{ddi}} \geq 0$ controls the penalty induced by harmful interactions with medications already retained in the prescription. At each step, we select $m^\star = \arg\max_{m\in\mathcal{C}_0\setminus\mathcal{R}} r_m(\mathcal{R})$. If $r_{m^\star}(\mathcal{R}) > 0$, we update $\mathcal{R} \leftarrow \mathcal{R}\cup\{m^\star\}$; otherwise, the procedure terminates. The final recommendation is given by $\widehat{y}_m = \mathbb{I}[m \in \mathcal{R}]$. This sequential refinement allows only medications that have already been retained to suppress conflicting candidates. Consequently, conflicting medications are not mutually penalized before either has been selected, favoring the candidate with stronger prescription support while progressively reducing the priority of incompatible alternatives. When $\lambda_{\mathrm{ddi}} = 0$, the procedure exactly recovers the original threshold-based candidate set $\mathcal{C}_0$. The refinement requires no additional training and is applied only during inference.

Table 1: Overall medication recommendation performance on eICU and MIMIC-IV. All metrics except Count MAE are reported in %. Results are reported as mean (std) over five random seeds. Best results are in **bold**, and second-best results are underlined.

	eICU					MIMIC-IV				
Model	Jaccard↑	F1↑	AP↑	C-MAE↓	DDI↓	Jaccard↑	F1↑	AP↑	C-MAE↓	DDI↓
LEAP	28.47 (0.31)	42.11 (0.35)	46.52 (0.38)	7.96 (0.13)	9.74 (0.16)	41.26 (0.27)	57.42 (0.29)	62.71 (0.33)	4.28 (0.09)	7.24 (0.13)
GAMENet	31.16 (0.24)	45.50 (0.28)	50.47 (0.31)	7.77 (0.11)	8.33 (0.14)	43.15 (0.22)	59.96 (0.25)	65.48 (0.29)	3.91 (0.08)	6.46 (0.11)
SafeDrug	31.72 (0.22)	46.13 (0.26)	51.24 (0.29)	7.29 (0.10)	7.62 (0.12)	43.91 (0.20)	60.39 (0.22)	66.28 (0.25)	3.96 (0.07)	5.76 (0.09)
4SDrug	30.24 (0.27)	44.02 (0.31)	48.63 (0.34)	8.95 (0.15)	5.73 (0.10)	43.42 (0.23)	60.12 (0.25)	65.83 (0.27)	4.63 (0.11)	4.42 (0.08)
COGNet	32.94 (0.20)	47.82 (0.23)	53.16 (0.26)	6.61 (0.08)	9.03 (0.15)	43.68 (0.21)	60.25 (0.23)	66.02 (0.26)	4.03 (0.08)	6.88 (0.11)
MoleRec	32.21 (0.20)	46.69 (0.23)	51.98 (0.27)	7.03 (0.09)	6.81 (0.11)	44.18 (0.19)	60.51 (0.21)	66.57 (0.23)	3.90 (0.07)	5.35 (0.08)
REFINE	33.37 (0.18)	48.44 (0.20)	54.11 (0.23)	5.83 (0.07)	6.25 (0.09)	44.46 (0.18)	60.60 (0.20)	66.86 (0.22)	3.86 (0.06)	4.61 (0.07)
SubRec	32.58 (0.19)	47.24 (0.22)	52.71 (0.25)	6.74 (0.08)	8.66 (0.13)	44.68 (0.17)	60.64 (0.19)	67.08 (0.21)	3.82 (0.06)	6.10 (0.09)
MGRec	33.26 (0.17)	48.21 (0.19)	53.72 (0.22)	6.02 (0.07)	7.14 (0.10)	44.82 (0.16)	60.58 (0.18)	67.29 (0.20)	3.79 (0.05)	4.98 (0.07)
CoREL	**35.77** (0.16)	**50.83** (0.18)	**56.90** (0.19)	**5.40** (0.06)	**4.94** (0.08)	**45.21** (0.14)	**60.81** (0.16)	**67.75** (0.18)	**3.75** (0.05)	**4.36** (0.06)

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP.

Datasets. We evaluate CoREL on three datasets spanning different clinical settings and patient populations. The eICU Pollard et al. (2018) and MIMIC-IV Johnson et al. (2023) are large-scale publicly available EHR databases. eICU serves as our primary multi-institutional benchmark, originally covering 208 healthcare institutions, with 135 retained after preprocessing, while MIMIC-IV provides a single-institution benchmark for conventional medication recommendation. We further include a real-world inpatient dataset from two tertiary hospitals in the same province in China. Medications in all datasets are mapped to DrugBank Wishart et al. (2018) identifiers, with DDI relations obtained from DrugBank v5.1. For each institution, patients are split into training, validation, and test sets at an 8:1:1 ratio, with all records of the same patient kept in the same split. Detailed preprocessing procedures and dataset statistics are provided in Appendix B.1.

Baselines. We compare our method with nine representative medication recommendation approaches: LEAP Zhang et al. (2017), GAMENet Shang et al. (2019), SafeDrug Yang et al. (2021), 4SDrug Tan et al. (2022), COGNet Wu et al. (2022), MoleRec Yang et al. (2023), REFINE Bhoi et al. (2023), SubRec Du et al. (2026), and MGRec Feng & Du (2026). Detailed descriptions of all baselines are provided in Appendix B.2.

Evaluation Metrics. We evaluate recommendation accuracy using Jaccard similarity, F1 score, and Average Precision (AP). We further report the DDI rate to assess population-level interaction safety and Count MAE to measure the discrepancy between predicted and observed prescription sizes. For multi-institutional datasets, all metrics are first computed within each institution and then macro-averaged across institutions, preventing larger institutions from dominating the evaluation. Formal definitions are provided in Appendix B.3.

Implementation Details. All baseline methods are implemented based on their original releases and re-tuned on the same validation splits for a fair comparison. For CoREL, all clinical embeddings are 128-dimensional. We fix the calibration shrinkage coefficient to $\lambda_b = 0.1$, select the global prediction threshold η by maximizing institution-macro F1 on the validation set, and choose λ_{ddi} to minimize validation DDI rate subject to an institution-macro F1 degradation of at most 0.005. Full hyperparameter settings and implementation details are provided in Appendix B.4.

5.2 OVERALL PERFORMANCE

Table 1 reports results on the multi-institutional eICU and single-institution MIMIC-IV benchmarks. CoREL achieves the best performance across all metrics, with the largest gains on eICU. Compared with REFINE, the strongest baseline in recommendation quality, CoREL improves Jaccard, F1, and AP by 2.40, 2.39, and 2.79 percentage points, respectively, while reducing Count MAE from 5.83 to 5.40 and DDI rate to 4.94%. The consistent gains in both set-based and ranking metrics indicate improved medication relevance estimation rather than a threshold-specific effect. On MIMIC-IV, CoREL remains competitive in the conventional single-institution setting and achieves the best

results across all five metrics. Table 2 further shows consistent improvements on the real-world two-center dataset, where CoREL achieves the best performance across all metrics, supporting its effectiveness under practical cross-institutional prescribing variation.

5.3 Validating Shared Patient-Dependent Evidence

Cross-Institutional Generalization of Patient Evidence. We evaluate whether the learned patient-dependent evidence generalizes to unseen institutions by splitting eICU into 96 training, 13 validation, and 26 held-out test institutions. No prescription labels from test institutions are used for learning or calibration. For each institution–medication pair, we measure whether the score ranks prescribed patients above unprescribed ones using $\mathrm{AUC}_{e,m} = \Pr(s_m(X^+) > s_m(X^-)) + \frac{1}{2}\Pr(s_m(X^+) = s_m(X^-))$. REFINE is evaluated using its pre-threshold continuous prediction scores on the same 23,192 test patients and institution–medication pairs. As shown in Figure 3(a), CoREL outperforms REFINE in all 26 held-out institutions, with an average AUC gain of 1.97 percentage points, indicating better cross-institutional generalization of patient-dependent evidence before local calibration.

Table 2: Results on the real-world two-center inpatient dataset. Metrics except Count MAE are reported in %. Results are mean (std) over five random seeds. Best results are in **bold**, and second-best results are underlined.

Model	Jaccard ↑	F1↑	AP↑	C-MAE↓	DDI↓
GAMENet	49.09 (0.28)	64.87 (0.31)	72.08 (0.36)	6.14 (0.11)	10.92 (0.14)
MoleRec	50.21 (0.22)	65.81 (0.24)	73.15 (0.27)	6.06 (0.08)	9.68 (0.11)
REFINE	<u>51.52</u> (0.19)	<u>67.05</u> (0.21)	<u>74.63</u> (0.24)	<u>5.95</u> (0.07)	9.18 (0.09)
MGRec	51.31 (0.18)	66.84 (0.20)	74.38 (0.22)	5.97 (0.06)	<u>9.52</u> (0.10)
CoREL	**52.71** (0.17)	**67.40** (0.19)	**74.95** (0.21)	**5.93** (0.06)	**8.41** (0.07)

Few-Shot Calibration for Unseen Institutions. Using the same held-out institutions, we split patients into an 80% support pool and a fixed 20% query set, and calibrate with 8, 16, 32, 64, or 128 support patients, averaging over 10 samplings. We compare CoREL with REFINE using either the same calibration or conventional fine-tuning under identical support sets. As shown in Figure 3(b), CoREL performs best across all support sizes and rapidly approaches its full-support F1 of 51.09% (694 support patients per institution on average). This shows that shared patient-dependent evidence enables sample-efficient adaptation to unseen institutions without gradient-based model updates.

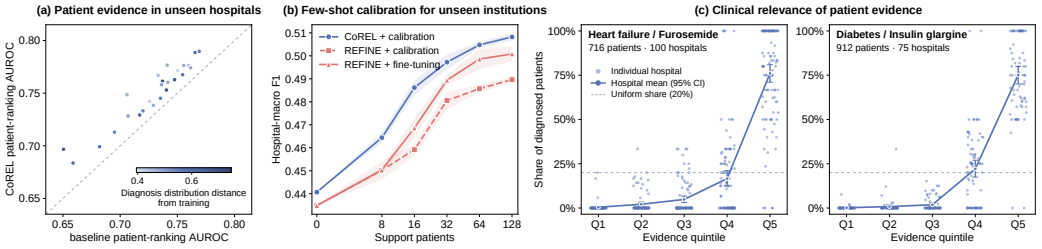

Figure 3: (a) Within-institution patient-ranking AUROC in unseen institutions, colored by diagnosis-distribution distance from training. (b) Recommendation F1 with increasing local support. (c) Distribution of patients with clinically related diagnoses.

Clinical Relevance of Patient Evidence. We assess whether the learned evidence aligns with established clinical knowledge using two diagnosis–medication associations: congestive heart failure–furosemide Heidenreich et al. (2022); McDonagh et al. (2023) and diabetes mellitus–insulin glargine Riddle et al. (2003). Within each eligible institution, patients are ranked by $s_m(x)$ and divided into five evidence quintiles. As shown in Figure 3(c), clinically related patients are strongly concentrated in Q5: 76.01% for heart failure–furosemide and 75.11% for diabetes–insulin glargine, compared with only 0.43% and 0.15% in Q1, respectively. This consistent concentration across institutions indicates that the learned evidence captures clinically meaningful patient–medication relationships rather than institution-specific prescribing frequency.

5.4 Institution-Specific Prescribing Tendency and Calibration

Systematic Residuals under Global Calibration. For each institution–medication pair, we compute the difference between the observed prescription rate and that predicted using the shared pa-

tient evidence with global tendency $\hat{b}_{0,m}$. Figure 4(a) shows substantial variation in both direction and magnitude across representative institution–medication pairs, indicating that global calibration leaves systematic local prescribing discrepancies.

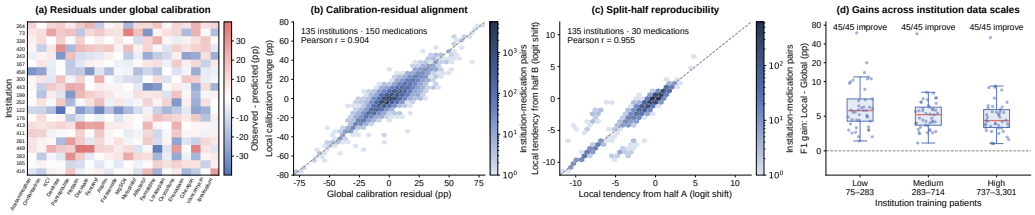

Figure 4: (a) Prescribing residuals under global calibration. (b) Alignment between institution-specific calibration and observed residuals. (c) Institution-specific tendencies estimated from disjoint halves of training patients. (d) Institution-level F1 gains across training-cohort-size tertiles.

Alignment between Local Calibration and Prescribing Residuals. We measure the local calibration correction as the change in mean predicted prescription probability from global to institution-specific calibration. As shown in Figure 4(b), these corrections closely align with the observed residuals ($r = 0.904$), and reduce the institution-averaged absolute prescribing-rate error by 58.4%, from 5.67 to 2.36 percentage points across all 135 institutions.

Reproducibility of Institution-Specific Tendencies. We split each institution's training patients into two disjoint halves and independently estimate local tendencies while keeping the evidence model and global reference fixed. The resulting estimates show strong agreement across 135 institutions and 30 frequent medications ($r = 0.955$; Figure 4(c)), supporting that the calibrated tendencies capture stable institution-specific prescribing patterns.

Calibration Gains across Institution Data Scales. We divide the 135 institutions into equal low-, medium-, and high-data groups. Local calibration improves F1 in all institutions, with median gains of 5.80, 5.22, and 4.36 percentage points, respectively (Figure 4(d)). The consistent gains across data scales indicate that local calibration is broadly effective, even with limited local observations.

5.5 Ablation Analysis

We ablate CREL (**C**), local calibration (**L**), DDI refinement (**D**), and shrinkage (**S**). **–C–L** uses pooled BCE with a global medication bias; **–C** uses BCE with institution-specific logits; **–L** uses the global tendency for all institutions; **–D** removes DDI-aware refinement; and **–S** removes shrinkage. Table 3 shows that removing CREL or local calibration substantially degrades recommendation quality, confirming the importance of separating shared evidence from institution-specific tendencies. Removing DDI refinement increases DDI from 4.94% to 8.20% and worsens C-MAE, while removing shrinkage causes a smaller but consistent degradation.

Table 3: Ablation results on eICU. Metrics except Count MAE are in %. Mean (std) over five random seeds.

Variant	Jaccard↑	F1↑	AP↑	C-MAE↓	DDI↓
–C–L	31.13 (0.24)	45.40 (0.27)	50.66 (0.30)	6.24 (0.09)	4.96 (0.09)
–C	33.58 (0.21)	48.02 (0.23)	54.29 (0.26)	6.08 (0.08)	5.24 (0.10)
–L	29.08 (0.26)	43.29 (0.29)	47.75 (0.33)	6.40 (0.10)	4.86 (0.09)
–D	36.21 (0.17)	51.02 (0.19)	57.11 (0.21)	7.12 (0.09)	8.20 (0.12)
–S	35.58 (0.18)	50.26 (0.20)	56.61 (0.22)	5.49 (0.07)	5.14 (0.09)
CoREL	35.77 (0.16)	50.83 (0.18)	56.90 (0.19)	5.40 (0.06)	4.94 (0.08)

6 Conclusion

We propose CoREL, a multi-institutional medication recommendation framework that separates shared patient-dependent prescribing evidence from institution-specific prescribing tendencies. By learning relative evidence through within-institution comparisons and recovering local prescribing behavior through hierarchical calibration, CoREL improves cross-institutional generalization while preserving institution-specific practice. Experiments on eICU, MIMIC-IV, and a real-world two-center dataset demonstrate consistent gains in recommendation accuracy and DDI safety, supporting a more structured approach to medication recommendation under institutional heterogeneity.

AI USE STATEMENT

Generative AI tools were used only for recommended-disclosure tasks, including language polishing, improving readability and manuscript organization, identifying relevant literature, refining scientific figures, and suggesting clearer presentation. They were not used to formulate or modify research hypotheses, design the core methodology or experiments, construct theoretical or conceptual frameworks, formulate mathematical claims or proofs, implement the proposed method, translate research content, preprocess or clean datasets, perform qualitative analysis, interpret experimental results, or generate synthetic data. All AI-assisted content was manually reviewed, verified, and revised by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This study uses retrospective electronic health record data. The public eICU and MIMIC-IV datasets were used under their respective access requirements, and the real-world two-center dataset was processed in accordance with applicable institutional and ethical requirements with identifying information removed. The model is intended for research rather than clinical use. Its recommendations may reflect biases in the underlying data and would require prospective validation and human oversight before deployment. Considering known DDIs does not guarantee patient-level clinical safety.

REPRODUCIBILITY STATEMENT

An anonymous repository linked in the abstract provides the core CoREL implementation, including model code, training pipelines, preprocessing scripts, and anonymized examples of the real-world data format. The paper and appendix further report model architecture, optimization, hyperparameters, preprocessing, evaluation, and experimental settings to support reproducibility.

REFERENCES

Julia Adler-Milstein and Ashish K Jha. Hitech act drove large gains in hospital electronic health record adoption. *Health affairs*, 36(8):1416–1422, 2017.

Nina L Barnett, Lelly Oboh, and Katie Smith. Patient-centred management of polypharmacy: a process for practice. *European Journal of Hospital Pharmacy*, 23(2):113–117, 2016.

Suman Bhoi, Mong Li Lee, Wynne Hsu, and Ngiap Chuan Tan. Refine: A fine-grained medication recommendation system using deep learning and personalized drug interaction modeling. *Advances in Neural Information Processing Systems*, 36:24013–24024, 2023.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

Wenjie Du, Xuqiang Li, Jinke Feng, Shuai Zhang, Wen Zhang, and Yang Wang. Integrating drug substructures and longitudinal electronic health records for personalized drug recommendation. *Advances in Neural Information Processing Systems*, 38:62950–62977, 2026.

Jinke Feng and Wenjie Du. Mgrec: Structure-grounded medication recommendation via condition-aware molecular representation learning. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 2599–2607. International Joint Conferences on Artificial Intelligence Organization, 8 2026.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Circulation*, 145 (18):e895–e1032, 2022.

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

Philip O Katz, Kerry Dunbar, Felice H Schnoll-Sussman, Katarina B Greer, Rena Yadlapati, and Stuart Jon Spechler. Acg clinical guideline: guidelines for the diagnosis and management of gastroesophageal reflux disease. *The American journal of gastroenterology*, 117(1):27, 2022.

Thomas A Lasko, Eric V Strobl, and William W Stead. Why do probabilistic clinical models fail to transport between sites. *NPJ Digital Medicine*, 7(1):53, 2024.

Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, et al. 2023 focused update of the 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. *European heart journal*, 44(37):3627–3639, 2023.

John M Neuhaus and Charles E McCulloch. Separating between-and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(5):859–872, 2006.

Aya F Ozaki, Cynthia A Jackevicius, Alice Chong, Maneesh Sud, Jiming Fang, Peter C Austin, and Dennis T Ko. Hospital-level variation in ticagrelor use in patients with acute coronary syndrome. *Journal of the American Heart Association*, 11(13):e024835, 2022.

Craig A Pedersen, Philip J Schneider, and Douglas J Scheckelhoff. Ashp national survey of pharmacy practice in hospital settings: prescribing and transcribing—2016. *American Journal of Health-System Pharmacy*, 74(17):1336–1352, 2017.

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

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Joel E Richter, Peter J Kahrilas, John Johanson, Paul Maton, Jeffrey R Breiter, Clara Hwang, Victoria Marino, Bernard Hamelin, Jeffrey G Levine, Esomeprazole Study Investigators, et al. Efficacy and safety of esomeprazole compared with omeprazole in gerd patients with erosive esophagitis: a randomized controlled trial. *The American journal of gastroenterology*, 96(3):656–665, 2001.

Matthew C Riddle, Julio Rosenstock, John Gerich, and Insulin Glargine 4002 Study Investigators. The treat-to-target trial: randomized addition of glargine or human nph insulin to oral therapy of type 2 diabetic patients. *Diabetes care*, 26(11):3080–3086, 2003.

Jill M Schachtner, Roy Guharoy, Joseph J Medicis, Nancy Newman, and Ronald Speizer. Prevalence and cost savings of therapeutic interchange among us hospitals. *American journal of health-system pharmacy*, 59(6):529–533, 2002.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1126–1133, 2019.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

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*, pp. 3970–3980, 2022.

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

Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. Conditional generation net for medication recommendation. In *Proceedings of the ACM web conference 2022*, pp. 935–945, 2022.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*, 2021.

Jenny Yang, Andrew AS Soltan, and David A Clifton. Machine learning generalizability across healthcare settings: insights from multi-site covid-19 screening. *NPJ digital medicine*, 5(1):69, 2022.

Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*, pp. 4075–4085, 2023.

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*, pp. 1315–1324, 2017.

A Derivation of the Conditional Relative Probability

We provide the derivation of Eq. equation 3. Consider two patient contexts x and x' observed at the same institution e for medication m. Let

$$p = p^{\star}_{e,m}(x), \qquad p' = p^{\star}_{e,m}(x'). \tag{12}$$

Assuming that the prescription outcomes of the two patients are conditionally independent given their clinical contexts and institution, we have

$$\Pr(Y_m = 1, Y'_m = 0 \mid X = x, X' = x', E = e) = p(1 - p'), \tag{13}$$

$$\Pr(Y_m = 0, Y'_m = 1 \mid X = x, X' = x', E = e) = (1 - p)p'. \tag{14}$$

Conditioning on the discordant event $Y_m + Y'_m = 1$, the probability that the patient with context x is the prescribed one is therefore

$$\begin{aligned} &\Pr(Y_m = 1 \mid Y_m + Y'_m = 1, X = x, X' = x', E = e) \\ &\quad = \frac{p(1 - p')}{p(1 - p') + (1 - p)p'}. \end{aligned} \tag{15}$$

From Eq. (1),

$$\frac{p}{1-p} = \exp\big(f_m^\star(x) + b_{e,m}^\star\big)\,, \qquad \frac{p'}{1-p'} = \exp\big(f_m^\star(x') + b_{e,m}^\star\big)\,. \tag{16}$$

Dividing the numerator and denominator of Eq. (15) by $(1-p)(1-p')$ gives

$$\begin{aligned}
&\frac{p/(1-p)}{p/(1-p) + p'/(1-p')} \\
&\quad = \frac{\exp\big(f_m^\star(x) + b_{e,m}^\star\big)}{\exp\big(f_m^\star(x) + b_{e,m}^\star\big) + \exp\big(f_m^\star(x') + b_{e,m}^\star\big)} \\
&\quad = \frac{\exp(f_m^\star(x))}{\exp(f_m^\star(x)) + \exp(f_m^\star(x'))} \\
&\quad = \sigma(f_m^\star(x) - f_m^\star(x'))\,.
\end{aligned} \tag{17}$$

Hence, conditioning on a discordant prescription pair removes the common institution-specific tendency $b_{e,m}^\star$, leaving a conditional probability determined solely by the relative patient-dependent prescribing evidence.

The formulation is symmetric with respect to the two patients. In particular, the probability that the patient with context x' is the prescribed one is

$$\sigma(f_m^\star(x') - f_m^\star(x)) = 1 - \sigma(f_m^\star(x) - f_m^\star(x'))\,. \tag{18}$$

B Experimental Setup Details

B.1 Dataset Statistics and Preprocessing

Table 4 summarizes the three datasets after preprocessing. They cover a multi-institutional ICU setting, a single-institution inpatient setting, and two tertiary hospitals in China, allowing us to evaluate medication recommendation under different institutional and clinical contexts.

Table 4: Statistics of the evaluation datasets after preprocessing. An encounter denotes an ICU unit stay in eICU and a hospital admission in MIMIC-IV and real-world dataset. Diag., Proc., and Med. denote diagnoses, procedures/interventions, and medications, respectively. For eICU, Proc. refers to retained non-drug treatment concepts. Vocabulary sizes are determined from training data. Encounter-level statistics count distinct retained labels in the current encounter, excluding historical context.

Item	**eICU**	**MIMIC-IV**	**Real-world**
# institutions	135	1	2
# patients	112,867	196,554	3025
# encounters	159,703	462,844	3748
Diag. vocabulary size	2,367	19,091	1471
Proc. vocabulary size	778	8,722	965
Med. vocabulary size	150	550	278
Avg./max encounters per patient	1.41 / 26	2.35 / 98	1.23/2
Avg./max diag. per encounter	3.46 / 118	12.76 / 56	3.41/20
Avg./max proc. per encounter	2.31 / 72	1.69 / 35	2.09/7
Avg./max med. per encounter	17.05 / 79	16.27 / 116	20.16/51
# training encounters	127,656	370,323	2998
# validation encounters	16,133	46,611	375
# test encounters	15,914	45,910	375

eICU serves as our primary multi-institutional benchmark. Its original records cover 208 hospitals, providing a setting for studying shared patient-dependent prescribing evidence alongside institution-specific prescribing tendencies. We retain adult ICU stays with at least one medication successfully

mapped to DrugBank, and subsequently select hospitals with at least 100 such stays. The resulting cohort contains 159,703 stays from 135 hospitals. Patient inputs include admission diagnoses, recorded diagnoses, medical history, non-drug treatment information, and demographic and admission characteristics. When available, completed prior ICU unit stays within the same hospitalization provide longitudinal clinical and medication context.

MIMIC-IV provides a complementary single-institution benchmark, allowing us to examine the method outside the multi-hospital setting. We use adult hospital admissions with recorded diagnoses and at least one retained medication label. Current-admission information consists of diagnosis and procedure codes together with patient and admission characteristics. Historical context includes diagnoses, procedures, and medications from the patient's earlier hospitalizations, restricted to admissions discharged before the current admission began. To limit extremely sparse prediction targets, we retain medications appearing in at least 100 eligible training admissions. The processed dataset contains 462,844 admissions and 550 medication labels.

The real-world inpatient dataset is collected from two tertiary hospitals in the same province in China. It complements the public datasets with records from a different healthcare system and provides an additional setting for examining institution-specific prescribing patterns.

Medication representation and preprocessing. We map medication records to DrugBank identifiers, using HICL-based ingredient mappings for eICU and NDC-based mappings for MIMIC-IV. Unresolved mappings are excluded, and repeated occurrences of the same medication within an encounter are merged into a single label. All medication labels observed in eICU training data are retained, whereas MIMIC-IV uses the frequency threshold described above. DDI resources are aligned with each dataset's medication vocabulary to construct a symmetric interaction matrix. A shared identifier system provides a consistent medication representation, while the retained vocabularies remain dataset-specific.

Task construction and data splitting. We construct one example $(x_q^{(t)}, e_q^{(t)}, y_q^{(t)})$ for each eligible encounter. The input comprises patient characteristics, current-encounter clinical information, and available clinical and medication records from preceding encounters, together with their temporal gaps. Current-encounter medications are excluded from the input and encoded as a multi-hot target vector over the dataset-specific medication vocabulary. Each example is associated with its institution identifier. Patients are assigned to training, validation, and test partitions with target proportions of 8:1:1, with all encounters of the same patient kept in one partition. Input and medication vocabularies are constructed exclusively from training data. During training, examples are grouped by institution to construct within-institution patient comparisons. Institution-specific calibration uses only the corresponding institution's training patients, while validation and test patients are reserved for model selection and final evaluation, respectively.

B.2 Baseline Descriptions

We summarize the main modeling strategy and DDI mitigation mechanism of each baseline below.

LEAP Zhang et al. (2017). LEAP formulates medication recommendation as a sequential generation problem, where an RNN decoder produces medications autoregressively. It further applies reinforcement-learning-based fine-tuning, using DDI-related penalties in the reward to discourage unsafe medication combinations.

GAMENet Shang et al. (2019). GAMENet combines longitudinal EHR representations with a graph-augmented memory module that incorporates both historical medication co-occurrence patterns and a global DDI graph. Medication safety is encouraged through an explicit DDI-aware regularization term during training.

SafeDrug Yang et al. (2021). SafeDrug enhances medication prediction with molecular structural information and DDI knowledge. It constructs complementary drug representations from molecular graphs and interaction structure, while penalizing unsafe co-prescriptions through a DDI-aware training objective.

4SDrug Tan et al. (2022). 4SDrug formulates medication recommendation from a set-to-set perspective, modeling both patient conditions and prescriptions as unordered sets rather than sequences. Its

prediction objective combines medication relevance with set-level safety regularization to reduce DDI risk.

COGNet Wu et al. (2022). COGNet adopts a Transformer-based autoregressive decoder with a copy-or-generate mechanism. At each decoding step, the model can reuse medications from historical prescriptions or generate new medications from the candidate vocabulary, thereby exploiting longitudinal treatment history. DDI-related information is incorporated into its medication representations and generation process.

MoleRec Yang et al. (2023). MoleRec models fine-grained associations between patient conditions and drug molecular substructures, allowing medication representations to be informed by both clinical context and molecular composition. For safety control, it dynamically adjusts the contribution of the DDI penalty according to the difference between the current rate and a predefined target.

REFINE Bhoi et al. (2023). REFINE extends conventional binary DDI modeling by explicitly accounting for interaction severity. It represents drug interactions using severity-aware weights and imposes stronger penalties on clinically more harmful co-prescriptions during optimization.

SubRec Du et al. (2026). SubRec integrates longitudinal EHR modeling with drug substructure representations to capture patient–drug associations at a finer molecular level. Its training objective additionally incorporates a DDI-related penalty to discourage unsafe medication combinations.

MGRec Feng & Du (2026). MGRec reframes medication recommendation as condition-aware inference in a continuous molecular representation space rather than direct prediction over discrete drug labels. It employs a Therapeutic–Safety Factorized Conditional Variational Autoencoder to separate treatment-relevant and safety-related molecular factors, and maps the inferred molecular representations back to clinically available medications. Safety information is incorporated through DDI-guided latent regularization, which aligns the safety subspace with the DDI graph and discourages high-risk co-recommendations.

B.3 Evaluation Metrics

For institution e, let

$$\mathcal{V}_e^{\text{test}} = \{(q,t) : e_q^{(t)} = e\}$$

denote its test prescription decisions, and let $\widehat{M}_q^{(t)} = \{m \in \mathcal{M} : \widehat{y}_{q,m}^{(t)} = 1\}$ denote the predicted medication set. The ground-truth prescription $M_q^{(t)}$ follows the notation in Section 3. We compute each metric within an institution and macro-average it across institutions.

Jaccard similarity. Jaccard similarity measures the overlap between the predicted and observed medication sets:

$$\text{Jaccard}_e = \frac{1}{|\mathcal{V}_e^{\text{test}}|} \sum_{(q,t)\in\mathcal{V}_e^{\text{test}}} \frac{|\widehat{M}_q^{(t)} \cap M_q^{(t)}|}{|\widehat{M}_q^{(t)} \cup M_q^{(t)}|}. \tag{19}$$

F1 score. The set-level F1 score jointly reflects precision and recall of the predicted prescription:

$$\text{F1}_e = \frac{1}{|\mathcal{V}_e^{\text{test}}|} \sum_{(q,t)\in\mathcal{V}_e^{\text{test}}} \frac{2|\widehat{M}_q^{(t)} \cap M_q^{(t)}|}{|\widehat{M}_q^{(t)}| + |M_q^{(t)}|}. \tag{20}$$

Average Precision (AP). AP evaluates the ranking quality of medication probabilities before thresholding. Let $\pi_q^{(t)}(k)$ denote the medication ranked at position k according to $\widehat{p}_{q,m}^{(t)}$, and define

$$P_q^{(t)}@k = \frac{1}{k} \sum_{j=1}^{k} \mathbb{I}[\pi_q^{(t)}(j) \in M_q^{(t)}]. \tag{21}$$

The average precision for a prescription decision is

$$\text{AP}_q^{(t)} = \frac{1}{|M_q^{(t)}|} \sum_{k=1}^{|\mathcal{M}|} P_q^{(t)}@k\, \mathbb{I}[\pi_q^{(t)}(k) \in M_q^{(t)}], \tag{22}$$

and the institution-level AP is

$$\mathrm{AP}_e = \frac{1}{|\mathcal{V}_e^{\mathrm{test}}|} \sum_{(q,t)\in\mathcal{V}_e^{\mathrm{test}}} \mathrm{AP}_q^{(t)} . \tag{23}$$

DDI rate. DDI rate measures the fraction of predicted medication pairs corresponding to known harmful interactions. Using the DDI matrix A defined in Section 3, we compute

$$\mathrm{DDI}_e = \frac{\sum_{(q,t)\in\mathcal{V}_e^{\mathrm{test}}} \sum_{1\le i<j\le|\mathcal{M}|} \widehat{y}_{q,i}^{(t)} \widehat{y}_{q,j}^{(t)} A_{i,j}}{\sum_{(q,t)\in\mathcal{V}_e^{\mathrm{test}}} \sum_{1\le i<j\le|\mathcal{M}|} \widehat{y}_{q,i}^{(t)} \widehat{y}_{q,j}^{(t)}}. \tag{24}$$

Count MAE. To evaluate prescription-size calibration, we measure the absolute difference between the number of predicted and observed medications:

$$\mathrm{CountMAE}_e = \frac{1}{|\mathcal{V}_e^{\mathrm{test}}|} \sum_{(q,t)\in\mathcal{V}_e^{\mathrm{test}}} \left| |\widehat{M}_q^{(t)}| - |M_q^{(t)}| \right| . \tag{25}$$

For any metric G, the reported multi-institutional result is

$$G_{\mathrm{macro}} = \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} G_e. \tag{26}$$

For the single-institution MIMIC-IV dataset, this reduces directly to the corresponding institution-level metric.

B.4 Implementation Details

Model architecture. Diagnosis, procedure, and medication codes are represented using 128-dimensional embeddings. Their pooled representations are concatenated at each visit, producing a 384-dimensional visit representation. Longitudinal patient history is encoded using a single-layer unidirectional GRU with an input dimension of 384 and a hidden dimension of 128. Patient-level attributes are projected to 128 dimensions and combined with the longitudinal representation through an MLP with dimensions $256 \to 256 \to 128$. We use GELU activations, dropout of 0.1, and LayerNorm in the representation network.

The patient-level attributes depend on data availability. For eICU, we include sex, ICU type, ICU admission source, hospital admission source, stay type, age, admission weight, and height. For MIMIC-IV, we include sex, admission type, admission source, and age. Missing categorical variables are represented using an explicit `missing` category. Age is normalized as $\min(\mathrm{age}, 90)/90$. For eICU, admission weight and height are scaled as $(w-80)/30$ and $(h-170)/20$, respectively, and clipped to $[-5, 5]$. Missing numerical values are set to zero together with additional indicators.

CREL optimization. The shared prescribing evidence model is optimized using AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, and batch size 128. Gradient norms are clipped at 1.0. Training proceeds for at most 50 epochs, with early-stopping patience of 8 epochs according to validation institution-macro F1. We do not use learning-rate scheduling and warm-up.

Each mini-batch contains records from a single institution. Training records are reshuffled at the beginning of every epoch, such that all training stays are covered once per epoch. For each medication, all valid discordant pairs within the mini-batch are used: each prescribed instance is paired with all unprescribed instances from different patients. Pairs formed by different stays of the same patient are excluded. Medications without both prescribed and unprescribed examples in a mini-batch do not contribute to the CREL loss, and a mini-batch is skipped if it contains no valid discordant pair.

Prescription calibration. After CREL training, the shared evidence model is fixed. For each medication, anchoring is performed over all training contexts by subtracting the mean shared prescribing score. The global and institution-specific prescribing tendencies are then estimated using a

safeguarded Newton solver. The global tendency is initialized at zero, while each institution-specific tendency is initialized from the corresponding global estimate. Newton optimization is run for at most 40 iterations and terminates when

$$\max_m \frac{|g_m|}{N} \leq 10^{-6}, \tag{27}$$

where g_m denotes the calibration gradient for medication m and N is the number of calibration observations. Calibration uses summed binary cross-entropy, and no gradient is propagated back to the shared evidence model. The institution-level shrinkage coefficient is fixed to $\lambda_b = 0.1$ and shared across institutions.

Prediction and DDI refinement. A single prediction threshold is shared across all medications and institutions. We search

$$\eta \in \{0.05, 0.075, \ldots, 0.95\} \tag{28}$$

and select the value maximizing validation institution-macro F1.

For DDI-aware refinement, we consider

$$\lambda_{\text{ddi}} \in \{0, 0.025, 0.05, 0.1, 0.2, 0.3, 0.5, 1\}. \tag{29}$$

We first retain configurations whose validation institution-macro F1 decreases by no more than 0.005 relative to the prediction without DDI refinement, and select the one with the lowest DDI rate. Remaining ties are resolved by higher F1 and then smaller λ_{ddi}. The refinement operates only on medications already included in the threshold-based candidate set and therefore never introduces medications whose original scores fall below the prediction threshold. When two candidates have identical selection scores, the medication with the smaller vocabulary index is selected first.

Experimental protocol. All institutions share the same model architecture and hyperparameters; no institution-specific hyperparameter tuning is performed. Validation data are used for checkpoint selection, early stopping, prediction-threshold selection, and selection of λ_{ddi}. Each experiment is repeated with five random seeds. Baselines are re-tuned on the corresponding validation splits under the same medication vocabulary, patient splits, and DDI matrix.

Computing environment. All models are implemented in Python 3.9.21 with PyTorch 2.7.0 and CUDA 11.8. Experiments are conducted on a single NVIDIA A100 GPU with 40 GB memory.

C Supplement Experiments

C.1 Computational Cost

Table 5: Computational cost comparison on eICU. Parameter counts are reported in thousands (K), training time in minutes, and inference time on the full test set in seconds.

Model	Params. (K) ↓	Train (min) ↓	Test (s) ↓
LEAP	488.31	76.84	3.974 ± 0.083
GAMENet	983.70	19.59	0.412 ± 0.025
SafeDrug	1318.44	29.84	1.684 ± 0.042
4SDrug	621.58	14.26	0.286 ± 0.018
COGNet	1054.82	63.17	4.862 ± 0.091
MoleRec	1846.27	34.72	1.927 ± 0.048
REFINE	1127.36	22.47	0.936 ± 0.031
SubRec	2148.63	48.52	2.741 ± 0.061
MGRec	1562.91	41.36	2.208 ± 0.054
CoREL	762.24	25.61	1.506 ± 0.038

We further compare the computational cost of different methods in terms of the number of trainable parameters, training time, and inference time. Table 5 reports the results under the same experimental environment. Training time measures the end-to-end training cost, while inference time denotes

the time required to complete prediction on the full test set. For inference time, we report the mean and standard deviation over 5 runs.

CoREL maintains a relatively compact model size, with 762.24K trainable parameters, smaller than most representation- and molecular-structure-based baselines. Its training time is moderate: although CREL requires within-institution pairwise comparisons and therefore incurs additional cost over simpler discriminative models such as GAMENet and REFINE, it remains substantially more efficient than more complex molecular and autoregressive approaches. At inference time, CoREL requires 1.506 seconds for the full test set. The additional inference cost mainly arises from sequential DDI-aware refinement, while institution-specific calibration introduces only lightweight scalar logit adjustments. Nevertheless, the overall inference cost remains considerably lower than that of autoregressive methods such as LEAP and COGNet.

Table 6: Illustrative eICU test cases selected to examine local calibration and DDI refinement. Green and red check marks denote selected medications present and absent in the recorded prescription, respectively; dashes denote non-selection. Only key medications are shown; F1 and DDI (%) are computed over complete prescriptions.

Case	Key medication	Recorded	REFINE	Global	Local	CoREL
Case A (Institution A): Sepsis; stage 3 chronic kidney disease; acute renal failure; metabolic acidosis; nausea/vomiting.						
A	Heparin	✓	✓	–	✓	✓
A	Ondansetron	–	✓	✓	✓	–
A	Pantoprazole	–	✓	✓	–	–
A	Docusate	–	–	–	✓	✓
A	Fentanyl	✓	–	✓	✓	✓
A	Magnesium sulfate	✓	–	✓	✓	✓
	F1 / DDI (%)	–	22.2 / 7.1	34.8 / 2.6	37.0 / 6.6	41.7 / 2.2
Case B (Institution B): Status post thrombectomy.						
B	Pantoprazole	–	✓	✓	–	–
B	Oxycodone	–	✓	✓	–	–
B	Acetylsalicylic acid	✓	✓	✓	✓	✓
B	Albuterol	–	✓	–	–	–
B	Famotidine	–	✓	✓	✓	–
B	Glucagon	–	✓	✓	–	–
	F1 / DDI (%)	–	54.0 / 8.4	59.3 / 6.7	67.9 / 8.7	63.8 / 2.4
Case C (Institution C): COPD; bone fractures; status post surgical fracture repair.						
C	Furosemide	–	✓	–	✓	–
C	Gabapentin	–	✓	✓	✓	–
C	Methylprednisolone	✓	✓	✓	✓	✓
C	Ondansetron	✓	✓	✓	✓	✓
C	Fentanyl	✓	✓	✓	✓	✓
C	Dextrose, unspecified form	–	✓	✓	–	–
	F1 / DDI (%)	–	36.0 / 8.6	40.9 / 6.7	45.7 / 6.9	45.2 / 3.3

C.2 Case Study

We examine three representative cases from the eICU test set to illustrate how institution-specific calibration and DDI-aware refinement affect medication prediction. The shared evidence model and all inference settings are fixed before case selection. Institution-specific tendencies are estimated exclusively from training patients, with no test labels used during calibration.

Table 6 compares REFINE with three stages of CoREL. *Global* combines the shared patient evidence with the global medication-specific tendency, while *Local* replaces it with the corresponding institution-specific tendency without changing the patient evidence. *CoREL* further applies DDI-aware refinement to the locally calibrated candidate set. Thus, the Global-to-Local comparison

isolates the effect of institution-specific calibration, whereas the Local-to-CoREL comparison reflects the effect of DDI-aware refinement. Green and red check marks denote selected medications that are present and absent in the recorded prescription, respectively. Only representative medication differences are shown, while F1 and DDI rates are computed from the complete predicted prescriptions.

In Case A, local calibration recovers heparin while removing pantoprazole, increasing F1 from 34.8% to 37.0%. Subsequent DDI refinement removes ondansetron and further improves F1 to 41.7%, while reducing the DDI rate from 6.6% to 2.2%. This case illustrates that local calibration and safety refinement can provide complementary improvements.

Case B highlights a clearer separation between the two components. Local calibration removes several medications not present in the recorded prescription, including pantoprazole, oxycodone, and glucagon, increasing F1 from 59.3% to 67.9%. DDI-aware refinement further reduces the DDI rate from 8.7% to 2.4%, although F1 decreases to 63.8%, illustrating the trade-off introduced by safety-oriented refinement.

A similar pattern appears in Case C. Local calibration removes dextrose but introduces furosemide, improving F1 from 40.9% to 45.7%. DDI refinement subsequently removes furosemide and gabapentin, substantially reducing the DDI rate from 6.9% to 3.3% while largely preserving recommendation accuracy (45.2% F1). Together, these examples illustrate how CoREL separates local prescribing adaptation from safety refinement, allowing each component to modify the prescription for a distinct purpose.

C.3 Hyperparameter Sensitivity

We analyze the sensitivity of CoREL to three key hyperparameters: the calibration shrinkage coefficient λ_b, the DDI penalty coefficient λ_{ddi}, and the representation dimension. Figure 5 reports validation performance under different settings.

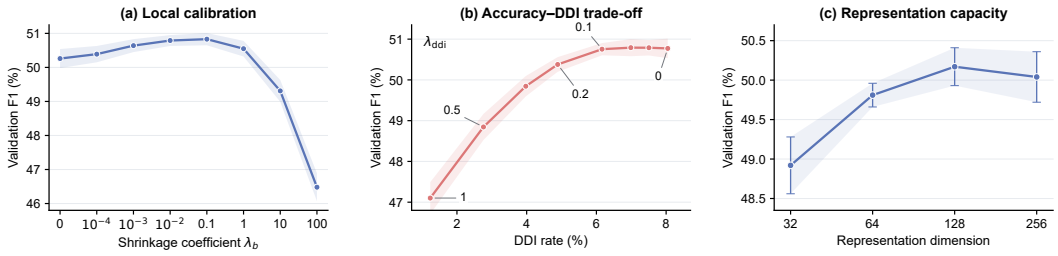

Figure 5: Hyperparameter sensitivity of CoREL on eICU.

Effect of calibration shrinkage. Figure 5(a) shows the effect of λ_b, which controls the strength of shrinkage from institution-specific tendencies toward the global prescribing tendency. Performance is relatively stable over a broad range of small to moderate values, with the best validation F1 achieved around $\lambda_b = 0.1$. When λ_b becomes too large, local tendencies are excessively constrained toward the global reference, substantially reducing the benefit of institution-specific calibration.

Accuracy–DDI trade-off. Figure 5(b) illustrates the trade-off induced by λ_{ddi}. Increasing λ_{ddi} progressively lowers the DDI rate, but eventually reduces validation F1 as stronger safety penalties suppress medications with high prescription support. Moderate values provide a favorable balance between recommendation accuracy and interaction safety, which motivates selecting λ_{ddi} on the validation set under the F1-degradation constraint described in Appendix B.4.

Effect of representation dimension. As shown in Figure 5(c), increasing the representation dimension from 32 to 128 consistently improves validation F1, indicating that additional capacity helps capture patient-dependent prescribing evidence. Increasing the dimension further to 256 yields little additional improvement, suggesting diminishing returns beyond 128 dimensions. We therefore use 128-dimensional representations in all main experiments.