Explainable Off-Policy Learning with Side Information for Intensive Care Unit Blood Test Orders

Anonymous Authors[1]

Abstract

Ordering a minimal subset of lab tests for patients in the intensive care unit (ICU) can be challenging. Care teams must balance ensuring the availability of the right information with reducing the clinical burden and costs associated with extracting and processing samples. In the current practice, most in-patient settings experience an over-ordering of lab tests with the intention of reducing the burden on the hospital and the environment. This paper leverages off-policy learning with side-information to identify the optimal set of ICU lab tests. Our approach facilitates the creation of assistive devices for clinicians to order lab tests by considering both the patient's observed status and their predicted future status. Incorporating clinically supported knowledge and using a support-guaranteed propensity score estimator for policy learning, our method demonstrates superior performance. The resulting application provides critical clinical information and reduces costs without omitting any vital lab orders, outperforming both a physician's policy and existing approaches to the problem.

1. Introduction

Laboratory tests are a key component in enabling doctors to make informed decisions regarding patient care. Clinicians order laboratory tests for diagnosing based on test results, determining necessary treatments for patients, and monitoring these results to observe improvements in patients' clinical conditions. The act of ordering laboratory tests is a crucial step in everyday medical practice.

However, laboratory tests often generate significant costs and most require blood draws or other invasive procedures, leading to increased blood loss, sleep disruption, and discomfort for patients. Moreover, the current medical practice tends to order redundant lab tests, significantly raising healthcare costs and ultimately becoming a burden on both patients and hospital resources (Feldman, 2009; Badrick, 2013). Research has shown that regular blood tests do not necessarily improve diagnoses (Iosfina et al., 2013; Pageler et al., 2013) and that ordering invasive laboratory tests can exacerbate a patient's condition (Berenholtz et al., 2004; Salisbury et al., 2011). In Intensive Care Units (ICU), where clinical teams face heightened challenges in determining necessary lab tests and patients are typically in critical conditions, the need for laboratory tests is often greater. Anecdotally, clinicians have observed that although ICU patients constitute only 5% of the hospital population, they account for 26% of the hospital's daily laboratory test volume. It has been noted that 20-40% of these laboratory tests could be eliminated without compromising patient safety (Zhi et al., 2013; Sedrak et al., 2016).

Given these factors, there is an urgent need for a clinical decision support tool that offers a second opinion on the optimal set of lab tests to order for each patient on a daily basis. Such a computer-aided policy could potentially reduce the cognitive load on clinicians, decrease costs, and improve hospital resource allocation. More importantly, it can improve patient conditions in the hospital. Motivated by these considerations we study how to create an explainable and reliable method for providing prescriptive advice on what lab tests to order.

We introduce a clinician-facing system tailored to the problem of ICU lab test ordering (See Figure 1). Our framework comprises three key steps. First, we learn a multi-variate, irregular time-series forecasting model to accurately predict a patient's future status. This model enables our policy to determine the necessary tests for each patient based on its predictability into the future as well its past trajectory. Next, we collect and incorporate expert level rules as side information, establishing bounds for the ideal lab test order at each time-step. This step is crucial, as the dataset of collected orders is unlikely to represent the optimal approach. Finally, our policy learns to prescribe the optimal set of lab tests for each patient based on maximizing the utility of the lab tests ordered by ensuring they adhere to clinical knowledge, and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

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

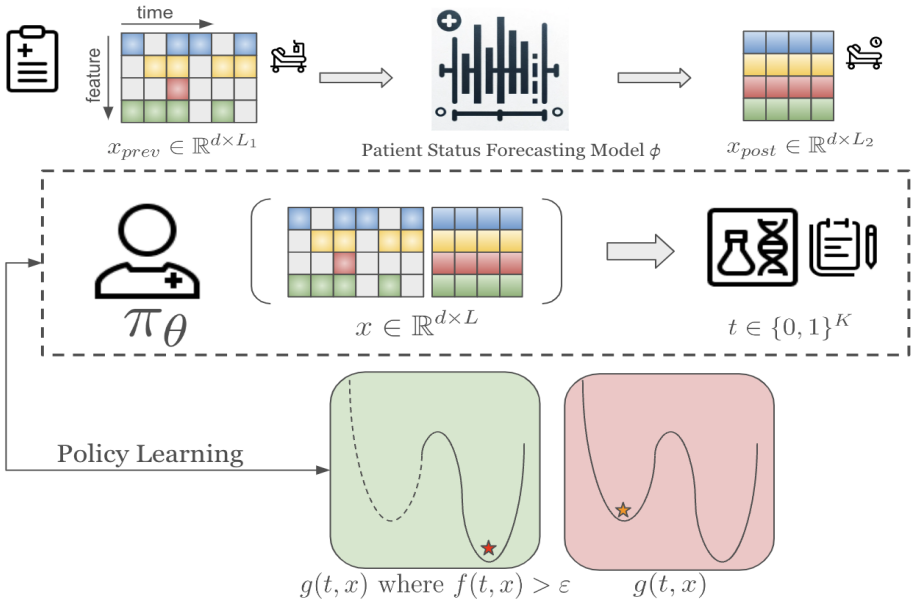

Figure 1. Overview of Proposed Method. Top: Development of an ICU patient status forecasting model ϕ for future predictions. Middle: Learning of policy π_θ to determine next-day lab test orders using observed and predicted patient data. Bottom: Implementation of a robust off-policy training algorithm to guarantee policy learning in areas with substantial support.

ensuring they minimize costs.

Contributions: (1) To the best of our knowledge, this is the first work to reorient policy learning objective towards creating a clinician-facing tool for ICU lab test ordering by leveraging clinical rules as side information. (2) Our method is inherently explainable. As clinicians receive lab test order suggestions, they also gain insight into the predicted values for each clinically relevant feature. This transparency naturally supports the decisions made by the learned policy. (3) Our off-policy learning algorithm introduces a novel metric to assess the usefulness of each lab test ordered. It simultaneously addresses the challenges of limited overlap and the irregularity of patient ICU stays as time-series data. Our extensive experiments conducted on real-world medical datasets showcase considerable promise for potential deployment as a clinical decision support tool.

2. Related Work

2.1. Multivariate Irregular Time-Series Forecasting in Healthcare

A variety of methods have been proposed for supervised learning on sparse, irregular, multivariate time-series data, especially concerning numeric EHR data. These approaches primarily utilize recurrent neural networks (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Che et al., 2018; Shukla & Marlin, 2021), with the increasing adoption of Transformer-based methods (Vaswani et al., 2017; Ren et al., 2021; Rasmy et al., 2021; Tipirneni & Reddy, 2022; Labach et al., 2023). While these methods are developed for medical event forecasting, they often fall short in forecasting future values. Meanwhile, the general domain of multivariate time-series value forecasting is well-explored (Wen et al., 2022; Wu et al., 2022a). Recent leading models are based on transformer architecture, such as Informer (Zhou et al., 2021a), FEDformer (Zhou et al., 2022), and PatchTST (Nie et al., 2022). PatchTST, in particular, segments the input series into patches for efficient processing and improved prediction accuracy. A notable lightweight MLP-based forecasting model, PatchTSMixer (Ekambaram et al., 2023), demonstrates performance comparable to PatchTST.

2.2. Off-Policy Learning with Constraints and Side Information

Off-policy learning refers to strategies where learning or decision-making processes are based on data independent of the current policy being improved (Lange et al., 2012; Sutton & Barto, 2018; Levine et al., 2020). Given that our approach utilizes previously collected ICU patient EHR data to learn a policy for lab test ordering, it aligns well with off-policy learning. However, off-policy learning, without further interaction with the environment, poses challenges, as the data-collection behavior policy may not be optimal. This issue is evident in lab test ordering, where clinicians often order redundant tests. Previous studies have shown that learned policies perform poorly with out-of-distribution actions (Fujimoto et al., 2019). To address this, some methods constrain the learned policy to actions within the support set of the behavior policy, either by introducing parameterized

generative models to estimate the behavior policy (Zhou et al., 2021b; Ghasemipour et al., 2021) or by incorporating regularization penalties to measure the divergence between the learned and behavior policies (Jaques et al., 2019; Kumar et al., 2019; Wu et al., 2019; 2022b; Mao et al., 2023). In the broader context of off-policy settings, Le et al. (2019) considered problems with multiple constraints, and Schweisthal et al. (2023) employed propensity score estimation to ensure non-trivial overlap. While extensive literature exists on ensuring the support of the behavior policy, the information is still extracted only from the observed dataset.

The wealth of medical knowledge accumulated from decades of critical care medicine provides valuable information for guiding policy learning. Some online bandit algorithms leverage side information (Slivkins, 2011; Kleinberg et al., 2019). In the off-policy setting, Wen et al. (2017) and Felicioni et al. (2022) have utilized action-related side information. To the best of our knowledge, there has not been a concrete, clinically actionable application of side information in the context of health. In this work, we showcase the impact that even simple rules summarized by clinicians can have as side information to guide policy learning.

2.3. Machine Learning for Laboratory Test Ordering

While there have been some rule-based attempts to reduce redundancy in laboratory test ordering, particularly in pediatric and cardiac surgical ICU settings ((Dewan et al., 2016; 2017; Kotecha et al., 2017)), these methods often lack generalizability to other contexts. Thus, there is a pressing need for data-driven approaches for the optimal scheduling of lab tests. Badrick (2013) employed a binary classifier to assess whether a given test contributes to information gain in the clinical management of patients with gastrointestinal bleeding. Similarly, Soleimani et al. (2017) used regression methods to model the novel information each test provides. However, these approaches are typically focused on specific diseases or patient types and do not generalize well across different patient cohorts. They also overlook a large number of clinical factors, such as predictive information from vital signs, the causal links between clinical interventions and test results, and the relative costs of ordering lab tests.

Off-policy learning, particularly in reinforcement learning, has been extensively explored for ICU patient treatment plans (Tang et al., 2022; Ma et al., 2023; Nambiar et al., 2023; Emerson et al., 2023; Kondrup et al., 2023; Schweisthal et al., 2023). However, only a few prior studies (Cheng et al., 2018; Chang et al., 2019) have delved into policy learning related to ICU patient measurements or lab tests. Cheng et al. (2018) focused on lab tests related to sepsis, examining only four lab features, whereas our method considers a broader cohort with 21 major blood work features. Chang et al. (2019) applied deep Q-learning to frequently occurring measurement features, including vital signs that do not require ordering, and did not make use of several important blood test features. Our method differs significantly in three key areas. We shift the focus of the device from optimizing solely for patient outcomes to also improving utility for clinicians; this change arose as an acknowledgement that clinicians are the primary end-users of such assistive software. Our learned policy accounts for both the patient's past observations and predicted future status. Finally, we explicitly handle the case where the behavioral policy exhibits limited support in the distribution over lab-tests enabling the reach of our method to a larger set of applications.

3. Notation

We model our problem as an off-policy **causal contextual bandits** problem with multi-dimensional binary actions. Given an observational dataset $\mathcal{D} = \{x_i, t_i, y_i\}_{i=1}^n$ with n i.i.d ICU patient stays [1]. $x_i \in X \subseteq \mathbb{R}^{d \times L}$ is the context, covariate, or a patient ICU stay represented by an irregular time-series matrix. $t_i \in T \subseteq \{0,1\}^K$ is the action, treatment, or the lab test order represented by a K-dimensional binary vector. $y_i \in Y \subseteq \mathbb{R}$ is the reward, outcome, or the utility of the lab test order represented by a real number.

Clinicians are keen to find a policy $\pi : X \to T$, which determines the lab tests to order for the following day given patient status as context (covariates). The policy value, $V(\pi) = \mathbb{E}_\pi[Y(\pi(X))]$, is the expected reward (outcome) of policy π. The function, $Y(\cdot) : T \to \mathbb{R}$, measures the potential outcome given actions (treatments) under our causal contextual bandits setting. Our objective is to find optimal policy π^* from policy class Π that maximizes the policy value $V(\pi)$, as expressed in the following equation:

$$\pi^* \in \arg\max_{\pi \in \Pi} V(\pi). \tag{1}$$

Since our problem is under the off-policy setting, $Y(\cdot)$ is not available. In our method, we use *conditional potential outcome function*, $g(t,x) = \mathbb{E}[Y(t) \mid X = x]$, to estimate individual potential outcome as a proxy of $Y(t)$. This makes our policy value, $V(\pi) = \frac{1}{n}\sum_{i=1}^n g(t_i, x_i)$, when evaluating the policy π on n samples. Specifically for the ICU blood test ordering problem, $g(t,x)$ is the *lab order usefulness function* and is predefined as a closed-formed function in our method. Apart from the causal setup, in addition to dataset $\mathcal{D}$, we utilize side information such as rules summarized by medical experts to help us mitigate the disparity between physician policy (i.e. logging policy) in real-world practice and our learned policy π. We assume that the logging policy is a sub-optimal policy in Π.

[1] Here each ICU stay i is considered a 'time step' under standard bandits setting.

Causal Assumptions: For the outcome estimation to be identifiable, we adhere to three standard assumptions in causal inference (Rubin, 1974): (1) *Consistency* ($Y = Y(T)$), asserting that observed outcomes align with potential outcomes under the observed treatment. (2) *Ignorability* ($Y(t) \perp\!\!\!\perp T \mid X \; \forall t \in \mathcal{T}$), confirming the absence of hidden confounders. (3) *Overlap* ($f(t,x) > \varepsilon, \forall x \in \mathcal{X}, t \in \mathcal{T}$, for some $\varepsilon \in [0, \infty)$), ensuring all potential treatments can be accurately estimated for every individual. Here, $f(t,x) = f_{T \mid X=x}(t)$ is the *global propensity score* (GPS) represents the conditional density of T given $X = x$ [2]. We differentiate between *weak* overlap ($\varepsilon = 0$) and *strong* overlap ($\varepsilon > 0$), focusing predominantly on *strong* overlap due to its enhanced reliability in finite sample contexts, unless specified otherwise.

Challenges: Restricted overlap introduces both empirical and theoretical hurdles. Firstly, datasets with high dimensionality or limited sample sizes often experience sparse coverage in the $X \times T$ space, leading to reduced overlap (D'Amour et al., 2021). This limitation increases uncertainty in our lab order usefulness function $g(t,x)$, hindering effective decision-making. Secondly, the possibility of small ε values in certain areas (due to unobserved patient trajectories or specific unassigned tests) necessitates a dependable off-policy approach.

Addressing Limited Overlap: Followed from Schweisthal et al. (2023), we confront this challenge by leveraging GPS. However, this is often not directly available and must be inferred from observational data. Our approach involves estimating the GPS as a probability density function $f(t,x) = f_{T \mid X=x}(t)$, correlating lab test orders T with the patient's ICU stay $X = x$. To achieve this, we opt for *conditional normalizing flows* (CNFs) (Trippe & Turner, 2018; Winkler et al., 2019) to estimate our GPS function. CNFs, built on the foundation of normalizing flows (Tabak & Vanden-Eijnden, 2010; Rezende & Mohamed, 2015), are fully-parametric generative models capable of modeling conditional densities $p(y \mid x)$. CNFs are able to transform a simple base density $p(z)$ through an invertible transformation, parameterized by $\gamma(x)$, dependent on the input x which makes CNFs suitable for density estimation. The training of CNFs is guided by minimizing the negative log-likelihood loss, $\mathcal{L}_{\text{nll}} = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}(t,x)$[3]. With our refined GPS function $\hat{f}(t,x)$, the policy training is steered away from areas of high uncertainty, enhancing the reliability of the resulting policy.

[2]GPS measures the probability of certain lab test t being ordered given patient ICU stay x.

[3]Additional details on the CNF training process are provided in the Appendix D.

4. Methodology

We introduce an approach to develop an optimal policy for ordering laboratory tests for ICU patients in a manner that is both explainable and dependable (refer to Figure 1). Initially, we construct a model to forecast patient conditions, which helps predict their future statuses. These predictions are then utilized as inputs for our policy formulation. We establish bounds for each observed lab test order based on clinically validated guidelines. Subsequently, we devise a potential outcome function to evaluate the effectiveness of each potential lab test order. Lastly, employing our off-policy learning algorithm, we devise an optimal and reliable policy for lab test ordering in ICU settings.

4.1. Explainable Context Setup

Previous studies (Cheng et al., 2018; Chang et al., 2019) typically represent patient covariates X as an imputed, irregular time-series matrix. Our interactions with medical experts revealed that clinicians order lab tests to gain updated insights into a patient's condition, aiming to provide superior care. They focus not only on current patient states but also on forecasting future developments, using subsequent lab results to validate their prognoses. This insight led us to include both observed and predicted patient states in our representation of patient covariates. Our forecasting model, ϕ, is based on a state-of-the-art transformer-based multivariate time-series forecasting model (Nie et al., 2022), taking as input the observed patient status X_{prev}. This input comprises a broad spectrum of features, including vital signs, treatments, and relevant lab test results, as recommended by ICU clinicians. The model aims to minimize the mean squared error ($\mathcal{L}_{mse}$) between the observed future status X^*_{post} and the predicted future status $X_{\text{post}} = \phi(X_{\text{prev}})$, using backpropagation for training. Further details about the construction of our patient status forecasting model are provided in Appendix B. By constructing the context X with patient past and predicted future status, we can naturally learn policies that are explainable. This setup can provide clinicians a better chance to evaluate policy actions during future deployments for 'online' approvals.

4.2. Clinically Supported Label Generation

Given that our dataset comprises observed treatments, the inferred physician policy may not represent the optimal approach. Nonetheless, lab test ordering in critical care is a well-established practice, underpinned by decades of clinical experience, which has yielded straightforward yet crucial guidelines for test ordering. For example, a clinician would typically order a Complete Blood Count (CBC) for a patient who has undergone a blood transfusion within the last 48 hours. One might question the necessity of policy learning when such guidelines exist. The answer

is twofold: (1) These guidelines are basic, derived from historical knowledge, and tend to be conservative, being triggered only when patients meet certain extreme criteria. (2) These guidelines apply to both past and future patient states, the latter of which clinicians cannot predict with precision during practice.

Such clinical insights provide a natural lower bound for lab test orders in our dataset. For each patient stay x, we can determine a lower bound $t^{lower} \in \{0,1\}^K$ for an order of K possible lab tests, based on these clinically derived guidelines applied to x. Considering our assumption that the observed treatment policy is suboptimal and possibly excessive, we establish the upper bound of each test order as $t^{upper} = \{t_j^* \vee t_j^{lower}\}_{j=1}^K \in \{0,1\}^K$, which is a combination of the guideline-derived order t^{lower} and the observed order t^*. This is because approximately 8%-12% of tests are missed in observed orders compared to guideline-based orders due to timing discrepancies or end-of-stay variations. The methodology for deriving these bounds is outlined in Algorithm 1. Given the conservative nature of the clinical guidelines, the guideline-generated orders constitute about 30% of the observed orders. Additional details on the rules and the test order bounds are available in Appendix C.

Algorithm 1 Find bound for observed lab test orders

Input: Patient stay x, Set of clinical rules $\mathcal{CR} = r^1, ..., r^M$
Output: Upper t^{upper} and lower t^{lower} bound lab test order of stay x
Determine observed test order t from x
$t^{upper}, t^{lower} \leftarrow \vec{0} \in \{0,1\}^K$ where K is number of lab tests
for r^m in $\mathcal{CR}$ **do**
  if x satisfies r^m **then**
    Find the indices $\mathcal{I} \subset \{1, ..., K\}$ of lab tests correspond to rule r^m
    $t_i^{lower} \leftarrow 1$ for $i \in \mathcal{I}$
  end if
  $t^{upper} \leftarrow \{t_j^* \vee t_j^{lower}\}_{j=1}^K$
end for

4.3. Potential Outcome Function for Lab Test Order Usefulness Assessment

Since clinicians planned (rules) and acted (observations) differently, our collected dataset $\mathcal{D}$ is not perfect, we mitigate this disparity by defining a expected outcome function $g(t,x)$ with multiple terms. This function is pivotal for assessing the utility of a lab test order t in the context of a patient's status x. It also serves as a critical estimator for the policy value $V(\pi)$. Unlike previous studies (Chang et al., 2019; Schweisthal et al., 2023) that primarily used mortality as a metric, our focus is on the usefulness of lab tests to clinicians rather than direct patient outcomes. This is due to the fact that lab tests principally aid clinicians in decision-making. Quantifying the exact usefulness of each lab test to clinicians is complex, but from our discussions with medical professionals, we identified key characteristics of an effective lab test order:

Informative: The lab tests should provide maximum information to clinicians. That is, if a test t_i predicts less variability than another test t_j, the necessity to order t_j becomes more significant. For a test order t and patient status $x = [x_{\text{prev}}, x_{\text{post}}]$ (comprising both observed past and predicted future statuses), the test result variation is defined as:

$$\Delta X(t,x) = \Delta_{avg}(t,x) + \Delta_{range}(t,x). \quad (2)$$

We calculate Δ_{avg} to gauge the mean variation of test values:

$$\Delta_{avg}(t,x) = \sum_{j=1}^{K} \mathbb{1}(t_j > 0.5) \cdot |\overline{x_{prev}^{lab,t_j}} - \overline{x_{post}^{lab,t_j}}|, \quad (3)$$

where $\overline{x^{lab,t_j}}$ signifies the average feature value corresponding to test t_j. The indicator term is used as $t \in [0,1]^K$ represents the predicted probability of each test being ordered. Δ_{range} measures the variation in the extremities of the test values ordered:

$$\Delta_{range}(t,x) = \sum_{j=1}^{K} \mathbb{1}(t_j > 0.5) \cdot \max(\delta_{max}, \delta_{min}), \quad (4)$$

where $\delta_{max} = |\max(x_{prev}^{lab,t_j}) - \max(x_{post}^{lab,t_j})|$ and

$$\delta_{min} = |\min(x_{prev}^{lab,t_j}) - \min(x_{post}^{lab,t_j})|$$

are the absolute differences between the maximum and minimum values of predicted and observed values for test t_j.

Within Bounds: Lab tests should conform to the bounds defined in Sec 4.2. The deviation of each test order from the bounds t_j^{lower} and t_j^{upper} is quantified by $\mathcal{L}_b$:

$$\mathcal{L}_b(t,x) = \sum_{j=1}^{K} \mathbb{1}(t_j^{upper} = t_j^{lower}) \cdot |t_j^{lower} - t_j|. \quad (5)$$

A lower value of $\mathcal{L}_b$ indicates that the test order t is more aligned with the specified bounds. Here we treat $t \in [0,1]^K$ as the predicted probability for each test for our policy.

Low Cost: The objective includes minimizing the cost of the ordered tests, aiming to reduce redundancy and, consequently, the financial and environmental burden. Differing from previous studies (Cheng et al., 2018; Chang et al., 2019) that assume uniform cost across tests, we consider the relative clinical costs of each test. The cost function is defined as:

$$C(t) = \sum_{j=1}^{K} \alpha_j \cdot \mathbb{1}(t_j > 0.5), \sum \alpha_j = 1, \quad (6)$$

where α_j represents the cost associated with each lab test.

Synthesizing these desirable qualities, we define the lab test order usefulness function (conditional outcome function) as:

$$g(t,x) = \Delta X(t,x) - \beta_1 \mathcal{L}_b(t,x) - \beta_2 C(t), \quad (7)$$

with β_1 and β_2 as regularization hyperparameters. This function, $g(t, x)$, is utilized to estimate the policy value $\hat{V}(\pi)$ in the subsequent policy learning phase.

4.4. Time-Aware Overlap-Guaranteed Policy Learning

Integrating the forecasting model, established bounds, and the potential outcome function, we introduce a time-aware, overlap-guaranteed off-policy learning algorithm. This algorithm is designed to create an explainable, reliable, and optimal policy for lab test ordering in ICU environments.

Our objective is to identify a policy π^{rel} that not only maximizes the estimated policy value $\hat{V}(\pi)$ but also guarantees that $V(\pi)$ is determined *reliably*. To this end, we restrict our policy search to regions within the covariate-treatment domain where data support is substantial, ensuring no violation of overlap. Modifying our original objective from Eq. (1), we reformulate it as:

$$\pi^{\text{rel}} \in \underset{\pi \in \Pi^{\text{r}}}{\arg\max}\ \hat{V}(\pi) \tag{8}$$

where $\Pi^{\text{r}} = \{\pi \in \Pi \mid f(\pi(x), x) > \bar{\varepsilon}),\ \forall x \in \mathcal{X}\}$ defines our policy class with a reliability threshold $\bar{\varepsilon}$, which dictates the minimum overlap. Given the constraints of our finite observational data, this leads to the following optimization problem:

$$\max_{\pi}\quad \frac{1}{n}\sum_{i=1}^{n} g\left(\pi(x_i), x_i\right) \quad \text{s.t.} \quad \hat{f}\left(\pi(x_i), x_i\right) \geq \bar{\varepsilon} \tag{9}$$

In this framework, the lab test order usefulness function $g(t, x)$ serves as our policy outcome estimator, and the GPS estimator $\hat{f}(t, x)$ limits the policy search space. To represent our policy π, we employ neural networks with learnable parameters π_θ. Since the constrained optimization problem in Eq.(9) is not amenable to direct learning through gradient updates, we convert it into an unconstrained Lagrangian problem:

$$\min_{\theta}\max_{\lambda_i \geq 0} -\frac{1}{n}\sum_{i=1}^{n}\left\{g\left(\pi_\theta(x_i), x_i\right) - \lambda_i\left[\hat{f}\left(\pi_\theta(x_i), x_i\right) - \bar{\varepsilon}\right]\right\}, \tag{10}$$

where $\pi_\theta(x_i)$ denotes the policy learner with parameters θ, and λ_i are the Lagrange multipliers for each sample i. This Lagrangian min-max objective is tackled through adversarial learning, employing gradient descent-ascent optimization techniques (Lin et al., 2020).

Leveraging our patient status forecasting model ϕ, the defined outcome estimation function g, the estimated GPS function $\hat{f}$, and the min-max-objective in Eq. (10), we are equipped to establish our explainable and reliable policy, as detailed in Algorithm 2. One important aspect to consider is that despite having defined our outcome estimation function g, it is imperative for all operations within g to be differentiable to enable the gradient descent-ascent algorithm to function effectively through backpropagation. In the case of C and ΔX, both employ a step function to ascertain which lab tests are ordered. We address this challenge by employing a modified Sigmoid function to approximate the step function operations.

Table 1. The test set performance for trained time-series forecasting model.

MODEL	MSE	MAE
LINEAR	0.037± 0.002	0.094± 0.005
LSTM	0.035± 0.004	0.072± 0.002
PATCHTSMIXER	0.032± 0.066	0.066± 0.003
PATCHTST	0.027± 0.001	0.059± 0.002

Algorithm 2 Reliable off-policy learning for ICU blood test ordering

input Data (X, T, Y), reliability threshold $\bar{\varepsilon}$
output Optimal reliable policy $\hat{\pi}_\theta^{\text{rel}}$
  // Step 1: Learn a multi-variate time-series forecasting model to predict future stay
  Estimate $\phi(x)$ that predicts patient next 24 hours based on the past 48 hours status
  // Step 2: Find lab order bounds and define lab test usefulness function $g(t, x)$
  Prepare t^{uppper}, t^{lower} and $g(t, x) = \Delta X(t, x) - \beta_1 \mathcal{L}_b(t, x) - \beta_2 C(t)$
  // Step 3: Estimate GPS using conditional normalizing flows
  Estimate $\hat{f}(t, x)$ via loss $\mathcal{L}_{\text{nll}}$
  // Step 4: Train policy network using reliable learning algorithm
  $\pi_\theta^{(k)} \leftarrow$ initialize randomly
  $\lambda \leftarrow$ initialize randomly
  for $m \in \{1, ..., M\}$ **do**
  for each epoch **do**
    for each batch **do**
      // Predict next 24 hours ICU stay with forecasting model
      $x_{post} \leftarrow \phi(x_{prev})$
      $x_i \leftarrow [x_{prev}, x_{post}]$
      $\mathcal{L}_\pi \leftarrow -\frac{1}{n}\sum_{i=1}^{n}\left\{g\left(\pi_\theta^{(m)}(x_i), x_i\right) - \lambda_i\left[\hat{f}\left(\pi_\theta^{(m)}(x_i), x_i\right) - \bar{\varepsilon}\right]\right\}$
      $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_\pi$
      $\lambda \leftarrow \lambda + \eta_\lambda \nabla_\lambda \mathcal{L}_\pi$
    end for
  end for
  end for
  // select best learned policy wrt constrained objective on validation set
  $\pi_\theta^{\text{rel}} \leftarrow \pi_\theta^{(m^*)}$, with $m^* =$
  $\arg\max_m \sum_{i=1}^{n} g\left(\pi_\theta^{(m)}(x_i), x_i\right) \cdot \mathbb{1}\left\{\hat{f}\left(\pi_\theta^{(m)}(x_i), x_i\right) > \bar{\varepsilon}\right\}$

5. Results

We conduct comprehensive experiments on a real-world ICU patient dataset to assess our proposed method.

MIMIC-IV: The MIMIC-IV database (Johnson et al., 2023) contains anonymized health records from patients in intensive care units. Our aim is to develop an optimal policy for daily lab test ordering in ICU patients, maximizing the utility of each test order (Y). In this context, every lab test order (T) is conceptualized as a K-dimensional binary vector. Based on clinical recommendations, we focus on $K = 10$ routinely conducted blood tests. We analyzed $n = 57,212$ valid patient ICU stays, each characterized by 71 irregular time-series features mirroring clinician daily practices. These features encompass lab test results, vital signs, and

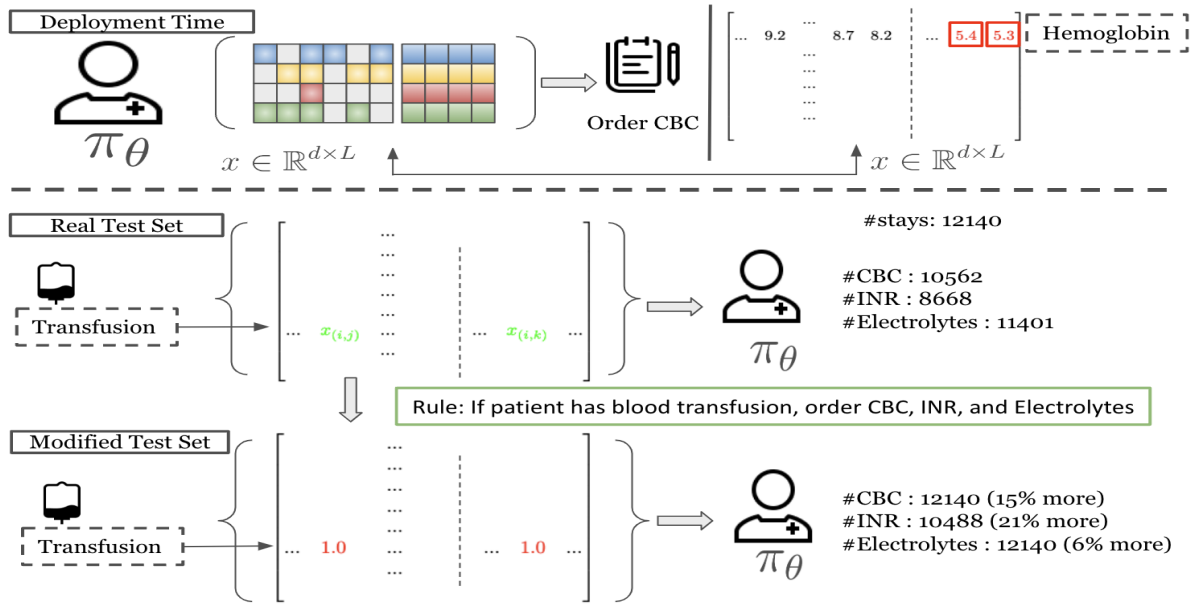

Figure 2. Integration of Medical Knowledge into Explainable Policy Learning. Top: At deployment, our policy transparently justifies actions, such as ordering a CBC test, based on predictions like a decrease in future Hemoglobin levels. Bottom: The policy incorporates clinical guidelines for lab test ordering, demonstrating adaptability by increasing specific test recommendations when the Transfusion feature value is altered.

Table 2. Testset performance for baseline and our learned policies.

POLICY	$\Delta X \uparrow$	COST$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
RANDOM$_{0.5}$	0.23	0.51	4.3	3.2	1.1
RANDOM$_{0.75}$	0.34	0.75	3.64	1.6	2.04
LOWERBOUND	0.37	0.62	0	0	0
UPPPERBOUND	0.44	0.82	0	0	0
PHYSICIAN	0.41	0.67	1.24	1.24	0
OURS(W/O GPS)	0.44	0.8	1.06	0.34	0.72
OURS(W GPS)	0.42	0.66	1.16	0.67	0.49

patient treatments. Further details on the preprocessing of the MIMIC-IV dataset are provided in Appendix A.

Patient Status Forecasting: To enable our policy to utilize both observed past and predicted future patient statuses, we initially train a multivariate time-series forecasting model based on observed ICU stays. The dataset, denoted as $\mathcal{D} = \{x^i_{prev}, x^{i*}_{post}\}_{i=1}^n$, consists of $x^i_{prev} \in \mathbb{R}^{48\times 71}$ representing the patient's past 48-hour ICU stay, and $x^{i*}_{prev} \in \mathbb{R}^{24\times 71}$ reflecting the *observed* true patient status for the subsequent day. We choose the transformer-based model PatchTST (Nie et al., 2022) for our patient status forecasting model ϕ due to its superior performance compared to other methods (see Table 1). The training of ϕ focuses on minimizing the mean squared error between the predicted future status x^i_{post} and the observed future [4].

Policy Training and Evaluation: We partitioned our dataset into training, validation, and test sets with proportions of 70%, 10%, and 20%, respectively. Initially, we utilize Algorithm 1 to establish order bounds for each patient stay. Subsequently, we train our estimated GPS function $\hat{f}(t, x)$, preserving the model parameters that has the lowest validation loss. We then employed Algorithm 2, executing $m = 5$ random restarts, to derive $\pi^{\text{rel}}_\theta : \mathbb{R}^{72\times 71} \to [0, 1]^{10}$ that yields the highest outcome on the GPS-constrained validation set. Further training specifics, including hyperparameters, are detailed in Appendix E.

For evaluation, we slightly modified our lab test order usefulness function $g(t, x)$ to suit as a metric. During testing, we set $\beta_1 = \beta_2 = 1$ for outcome calculation in Eq. (7). Additionally, we adapted the smooth term $\mathcal{L}_b$ from Eq. (5) to a discrete form:

$$\mathcal{L}_b^{test}(t, x) = L_{low} + L_{up},$$
$$L_{low} = \sum_{j=1}^{K} \mathbb{1}(t_j < 0.5, t_j^{low} = 1) \text{ and}$$
$$L_{up} = \sum_{j=1}^{K} \mathbb{1}(t_j > 0.5, t_j^{up} = 0),$$

where L_{up} indicates the redundant tests ordered, and L_{low} represents the essential tests missed by t. ΔX quantifies the variability (information) of the clinician's test order t, while $C(t, x)$ denotes the actual lab test cost.

Baselines: To date, no studies have directly focused on

[4] Additional details on the training of our forecasting model are available in Appendix B.

Table 3. Testset performance of prior work and ours policies.

POLICY	$\Delta X \uparrow$	COST$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	INFO GAIN$\uparrow$
PHYSICIAN	0.41	0.67	1.24	0.99
RL (LOW COST)	-	0.62	1.8	1
RL (HIGH COST)	-	0.8	2.3	1.4
OURS(W/O GPS)	0.44	0.8	1.06	(1.2)
OURS(W GPS)	0.42	0.66	1.16	(0.98)

explainable and reliable off-policy learning for ICU lab test ordering. However, as each lab test order t is a binary vector, we compared our trained policies against random policies, lower and upper bound policies, and, crucially, the observed clinician policy. In Table 2, we evaluate our baseline models, including random policies with 50% and 75% probabilities of ordering a test. With increasing orders and costs, the variability provided by these policies also rises. Nevertheless, random policies, despite higher costs, yield no substantial information. From lower bound to physician, and then to upper bound policy, we observe a trend of increasing test order information and cost. Since bound policies represent the limits of our lab test order space, they exhibit a bound metric of 0. The physician policy, not entirely aligned with clinical rule-generated orders, incurs some L_{lower}. Our goal is to discover policies with higher ΔX and lower out-of-bound test orders $\mathcal{L}_b$ at minimal cost.

GPS for Reliable Policy Learning: Our methodology ensures the discovery of reliable policies based on the estimated GPS function $\hat{f}(x,t)$. We set the reliability threshold $\bar{\varepsilon}$ at the 5%-quantile of all $\hat{f}(x,t)$ in the training dataset. As shown in Table 2, the average total outcome of a reliable policy is approximately 12% higher than that of a policy trained without GPS constraints. Notably, both our reliable and naively trained policies surpass the Physician policy by maintaining lab test orders within bounds or reducing costs, all while providing high information value to clinicians.

Comparing with RL Policy: Although our approach to lab test ordering is framed within a general off-policy setting via a causal inference perspective, rather than as a Markov Decision Process (MDP), we find it instructive to compare our results with the work of Chang et al. (2019). Their approach conceptualizes measurement scheduling as a deep Q-learning task in an offline-RL setting. We adapted their method to our dataset, treating each lab test order as an action. While their methodology treats patient status as an hidden state vector of a mortality classifier, making the evaluation of ΔX intractable, we can still compare policies based on $\mathcal{L}_b$ and cost. We observed that a reward system solely based on a single value (mortality) is inadequate for the lab test ordering problem, as the learned policy does not directly benefit the patient. In Table 3, RL policies derived from Chang et al. (2019)'s method tend to have either low cost with a high $\mathcal{L}_b$ or high cost with significant out-of-bound orders. Interestingly, under their policy evaluation framework, which calculates cumulative information gain from the mortality classifier, our methods demonstrate comparable performance.

Policy Explainability and Rule Learning: Figure 2 demonstrates the explainability of our policy, illustrating how lab test orders are linked to both past observations and future predictions of patient status, with each recommendation supported by a patient status time-series matrix for clear clinical rationale. Our experiments further validate the policy's ability to adhere to basic clinical guidelines by showing that adjustments in the Blood Transfusion variable or extreme changes in predicted lab values lead to an increase in specific test orders. These findings, detailed in Appendix G, underscore the policy's capacity to integrate critical clinical insights, enhancing its applicability and trustworthiness in a healthcare setting.

6. Discussion and Future Work

In this paper, we introduce a novel approach for learning optimal lab test ordering policies for ICU patients, focusing on reliability and explainability. Our method addresses the limited overlap in treatment and covariate spaces; however, challenges such as ignorability may persist due to potential unobserved confounders or the presence of corrupted and noisy data. A critical element of our approach is the accurate forecasting of patient future status. We adopted a general time-series forecasting method from Nie et al. (2022), but ICU data often exhibits strong correlations and underlying structures that could be better captured using a medical knowledge graph. Future improvements might include integrating graph-based time-series models to enhance the accuracy and interpretability of patient status predictions.

Our current framework utilizes a general off-policy learning and causal setup. However, the outcome function $g(t,x)$ could also serve as a reward function in an offline reinforcement learning (RL) context, provided state and status representations are adequately defined. We demonstrate that using mortality as a sole guide is insufficient for lab test ordering, but offline RL remains a viable avenue for exploration in this domain. This study does not account for potential distribution shifts in trained policies, as our data originates from a single source. When applying these policies to different patient cohorts, there is a risk of catastrophic forgetting. Therefore, developing methods that can adapt to distribution shifts is crucial.

Our methodology paves the way for deploying reliable and explainable policies in real clinical settings, with the potential for real-time feedback from medical professionals. Data collected from such deployments would be invaluable for refining policy accuracy through ground truth labeling and counterfactual learning.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. This work presents a system to prioritize which lab-tests are prescribed for a patient. Prior to deployment, there is a need to assess the generalizability of the resultant model on varied healthcare datasets as well as perform a careful study to ensure that the model does not learn or exacerbate biases in clinical practice.

References

Badrick, T. Evidence-based laboratory medicine. *The Clinical Biochemist Reviews*, 34(2):43, 2013.

Berenholtz, S. M., Pronovost, P. J., Lipsett, P. A., Hobson, D., Earsing, K., Farley, J. E., Milanovich, S., Garrett-Mayer, E., Winters, B. D., Rubin, H. R., et al. Eliminating catheter-related bloodstream infections in the intensive care unit. *Critical care medicine*, 32(10):2014–2020, 2004.

Chang, C.-H., Mai, M., and Goldenberg, A. Dynamic measurement scheduling for event forecasting using deep rl. In *International Conference on Machine Learning*, pp. 951–960. PMLR, 2019.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1), 2018.

Cheng, L.-F., Prasad, N., and Engelhardt, B. E. An optimal policy for patient laboratory tests in intensive care units. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pp. 320–331. World Scientific, 2018.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Dewan, M., Gálvez, J. A., Polsky, T., Kreher, G., Kraus, B., Ahumada, L. M., McCloskey, J. J., and Wolfe, H. Reducing unnecessary postoperative complete blood count testing in the pediatric intensive care unit. *The Permanente Journal*, 2016.

Dewan, M., Galvez, J., Polsky, T., Kreher, G., Kraus, B., Ahumada, L., McCloskey, J., and Wolfe, H. Reducing unnecessary postoperative complete blood count testing in the pediatric intensive care unit. *The Permanente journal*, 21, 2017.

Dolatabadi, H. M., Erfani, S., and Leckie, C. Invertible generative modeling using linear rational splines. In *AISTATS*, 2020.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In *NeurIPS*, 2019.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Ekambaram, V., Jati, A., Nguyen, N., Sinthong, P., and Kalagnanam, J. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. *arXiv preprint arXiv:2306.09364*, 2023.

Emerson, H., Guy, M., and McConville, R. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. *Journal of Biomedical Informatics*, 142:104376, 2023.

Feldman, L. Managing the cost of diagnosis. *Manag Care*, 5:43–45, 2009.

Felicioni, N., Ferrari Dacrema, M., Restelli, M., and Cremonesi, P. Off-policy evaluation with deficient support using side information. *Advances in Neural Information Processing Systems*, 35:30250–30264, 2022.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Ghasemipour, S. K. S., Schuurmans, D., and Gu, S. S. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *International Conference on Machine Learning*, pp. 3682–3691. PMLR, 2021.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Iosfina, I., Merkeley, H., Cessford, T., Geller, G., Amiri, N., Baradaran, N., Norena, M., Ayas, N., and Dodek, P. M. Implementation of an on-demand strategy for routine blood testing in icu patients. In *D23. QUALITY IMPROVEMENT IN CRITICAL CARE*, pp. A5322–A5322. American Thoracic Society, 2013.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66 (4):1–77, 2019.

Kondrup, F., Jiralerspong, T., Lau, E., de Lara, N., Shkrob, J., Tran, M. D., Precup, D., and Basu, S. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15696–15702, 2023.

Kotecha, N., Shapiro, J. M., Cardasis, J., and Narayanswami, G. Reducing unnecessary laboratory testing in the medical icu. *The American journal of medicine*, 130 (6):648–651, 2017.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Labach, A., Pokhrel, A., Huang, X. S., Zuberi, S., Yi, S. E., Volkovs, M., Poutanen, T., and Krishnan, R. G. Duett: Dual event time transformer for electronic health records. *arXiv preprint arXiv:2304.13017*, 2023.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pp. 45–73. Springer, 2012.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020.

Ma, H., Zeng, D., and Liu, Y. Learning optimal group-structured individualized treatment rules with many treatments. *Journal of Machine Learning Research*, 24(102): 1–48, 2023.

Mao, Y., Zhang, H., Chen, C., Xu, Y., and Ji, X. Supported trust region optimization for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 23829–23851. PMLR, 2023.

Melnychuk, V., Frauen, D., and Feuerriegel, S. Normalizing flows for interventional density estimation. In *ICML*, 2023.

Nambiar, M., Ghosh, S., Ong, P., Chan, Y. E., Bee, Y. M., and Krishnaswamy, P. Deep offline reinforcement learning for real-world treatment optimization applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4673–4684, 2023.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Pageler, N. M., Franzon, D., Longhurst, C. A., Wood, M., Shin, A. Y., Adams, E. S., Widen, E., and Cornfield, D. N. Embedding time-limited laboratory orders within computerized provider order entry reduces laboratory utilization. *Pediatric Critical Care Medicine*, 14(4):413–419, 2013.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13, 2021.

Ren, H., Wang, J., Zhao, W. X., and Wu, N. RAPT: Pre-training of time-aware transformer for learning robust healthcare representation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Salisbury, A. C., Reid, K. J., Alexander, K. P., Masoudi, F. A., Lai, S.-M., Chan, P. S., Bach, R. G., Wang, T. Y., Spertus, J. A., and Kosiborod, M. Diagnostic blood loss from phlebotomy and hospital-acquired anemia during acute myocardial infarction. *Archives of internal medicine*, 171(18):1646–1653, 2011.

Schweisthal, J., Frauen, D., Melnychuk, V., and Feuerriegel, S. Reliable off-policy learning for dosage combinations. *arXiv preprint arXiv:2305.19742*, 2023.

Sedrak, M. S., Patel, M. S., Ziemba, J. B., Murray, D., Kim, E. J., Dine, C. J., and Myers, J. S. Residents' self-report on why they order perceived unnecessary inpatient

laboratory tests. *Journal of hospital medicine*, 11(12): 869–872, 2016.

Shukla, S. N. and Marlin, B. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Slivkins, A. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 679–702. JMLR Workshop and Conference Proceedings, 2011.

Soleimani, H., Hensman, J., and Saria, S. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1948–1963, 2017.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tabak, E. G. and Vanden-Eijnden, E. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Tang, S., Makar, M., Sjoding, M., Doshi-Velez, F., and Wiens, J. Leveraging factored action spaces for efficient offline reinforcement learning in healthcare. *Advances in Neural Information Processing Systems*, 35:34272–34286, 2022.

Tipirneni, S. and Reddy, C. K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2022.

Trippe, B. L. and Turner, R. E. Conditional density estimation with Bayesian normalising flows. *arXiv preprint arXiv:1802.04908*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wen, M., Papusha, I., and Topcu, U. Learning from demonstrations with high-level side information. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

Winkler, C., Worrall, D., Hoogeboom, E., and Welling, M. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022a.

Wu, J., Wu, H., Qiu, Z., Wang, J., and Long, M. Supported policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 31278–31291, 2022b.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Zhi, M., Ding, E. L., Theisen-Toupal, J., Whelan, J., and Arnaout, R. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PloS one*, 8(11):e78962, 2013.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021a.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pp. 27268–27286. PMLR, 2022.

Zhou, W., Bajracharya, S., and Held, D. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pp. 1719–1735. PMLR, 2021b.

A. Datasets and Data preprocessing

MIMIC (Medical Information Mart for Intensive Care) is a publicly available database of de-identified electronic health records (EHRs) from patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. MIMIC-IV (Johnson et al., 2023) is one of the largest and most comprehensive critical care databases available, containing data from over 300,000 hospital admissions between 2008 and 2019.

The MIMIC-IV dataset includes a wide range of clinical data, such as vital signs, laboratory test results, medication orders, procedures, diagnoses, and demographic information. The data is collected from various sources, including bedside monitors, electronic medical records, and nursing notes, among others. The data is stored in a relational database format, with each record corresponding to a specific patient encounter. To ensure patient privacy and confidentiality, the MIMIC-IV dataset is de-identified and follows the Health Insurance Portability and Accountability Act (HIPPA). It is released under a data use agreement, which requires users to follow strict guidelines for data security and ethical use. However, access to the dataset is free for researchers and clinicians who agree to these terms. Overall, the MIMIC-IV dataset is a valuable resource for developing and testing predictive models, evaluating interventions, and improving ICU patient outcomes. With the success of its predecessor, the MIMIC-IV dataset was just released and has not been fully explored.

Preprocessing Pipeline for MIMIC-IV dataset

We develop a set of Python scripts that preprocess and aggregate the MIMIC-IV raw data from relational database format into a format that can be utilized by deep-learning community. For developing our preprocessing procedure, we followed and extended a prior work bench-marking the MIMIC-III (Johnson et al., 2016) with Python (Harutyunyan et al., 2019).

We first create a folder indexed by patient subject identification number and extract each patient's raw admission, ICU stays, diagnoses, and laboratory, input/output events information and saved into each patient folder. We then validate the extracted value and unify the missing values obtained from the raw data for each patient. After this step, we prepare the patient ICU stay into time-series data with episodes by event time stamps and store each episode's outcome (mortality, length of stay, diagnoses) in a separate file. To reproduce the work done by (Chang et al., 2019), we also generate a script to convert each patient's diagnoses codes into a multi-hot time-invariant features.

Sitting down with clinical experts in ICU department, we hand-picked the relevant features that clinicians would consider during their daily practice. The features that we considered were Hemoglobin, White cells/White blood cell count, Platelets, Sodium (Na), Potassium (K), Calcium, Phosphate, Magnesium, INR (PT/INR), Alkaline phosphatase (ALP), Bilirubin, ALT, Lactate, Partial pressure of carbon dioxide/PaCO2, PaO2, pH, Bicarb/Bicarbonate, Blood urea nitrogen, Creatinine (blood), Troponin, Creatinine phosphokinase (kinase), Diastolic blood pressure, Mean blood pressure, Systolic blood pressure, Temperature, Heart Rate, Arterial Blood Pressure mean (ABPm), Urine Output, Fluid balance, Fraction inspired oxygen (FiO2), RR Respiratory Rate, Ventilation (mode) Ventilation Mode, PEEP Positive End-Expiratory Pressure, Vt Tidal Volume, GCS Glasgow Coma Scale, SAS Richmond Agitation-Sedation (RAS) Scale, ICDSC Intensive Care Delirium Screening Checklist, Sedation (infusions), Analgesia (infusions), Antipsychotics, Dialysis (yes/no), Vasopressors (IV/PO), Dialysis (output), TPN, Transfusions of blood products, liver toxic drug, Antibiotics, Prone position, NO, Paralysis, Steroids, Diuretics, Antihypertensives (IV/PO), Anticoagulants, Antiepileptics, Enteral nutrition, PPI, Antiarrhythmics, Xray, US, MRI, CT scans, EKG, EEG, ECHO, Hepatotoxic drugs. However, MIMIC-IV data doesn't have any occurrences or record of certain features (e.g. Ultra Sound or NO), we finally picked 71 features with some merging of features with different code like Temperature (°F) and Temperature (°C) and some non-merged feature like Vasopressors. We show the occurrences of features we considered in Table A.

Finally, we convert each patient stay episodes into a patient status forecasting dataset $\mathcal{D} = \{X_i\}_i^N$ where X represents a irregular time-series matrix of patient stay i. Among the features in Table A, the first 21 features are the test result values correlated with 10 common blood test we consider to order/not order in this paper.

The labels of the dataset are indicators of whether the patient passed away after their ICU stay. In order to perform our irregular time-series patient mortality classification, we have to check whether each ICU stay's end time is before the record time of death of the patients. In order for our model to learn meaningful representation, we also eliminated the ICU stays with duration less than 12 hours and stays that has less than 5 lab tests ordered.

After preprocessing with these basic criterion of the ICU stays, we selected 57,212 ICU stays. The morality rate of the total stays is around 12%.

Table 4. Occurrences of selected feature on MIMIC-IV dataset

Measurement	Occurrence	Measurement	Occurrence
Hemoglobin	272244	Heart Rate	4595306
WBC	260106	Urine Output	2381884
Platelets	261397	Respiratory Rate	4549152
Sodium	289172	Ventilation	443722
Potassium	360470	PEEP	433118
Calcium	267206	Tidal Volume	411535
Phosphate	265319	GCS	3468710
Magnesium	287194	SAS	208084
INR	183070	ICDSC	207757
ALP	72296	Sedation	416082
Bilirubin	73361	Propofol	281361
ALT	73237	Analgesia	346169
Lactate	156908	Antipsychotics	9472
PaCO2	256789	Dialysis	242060
PaO2	218941	Vasopressors	439452
ph	258136	TPN	5655
Bicarbonate	285777	Transfusions	54684
Creatinine	291750	Prone Position	2859
Blood Urea Nitrogen	295945	Paralysis	12179
Troponin	28772	Diuretics	81959
Creatinine Kinase	35863	Antihypertensives	199045
Diastolic Blood Pressure	4504384	Anticoagulants	88757
Mean Blood Pressure	4507939	Antiepileptics	18394
Systolic Blood Pressure	4510870	Enteral Nutrition	8939
Temperature	1068275	PPI	89363
FiO2	571814	Antiarrhythmics	14216

B. Details of Building Multivariate Time-Series Patient Status Forecasting Model

In our study, we developed a multivariate time-series model to forecast patient status, leveraging deep learning techniques for both short and long-term predictions. We evaluated various architectures including simple linear transformations, Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), PatchTSMixer (Ekambaram et al., 2023), and PatchTST (Nie et al., 2022), focusing on their ability to accurately predict future patient states based on historical data.

PatchTST stands out for its innovative approach to handling time-series data, treating inputs and outputs as matrices to effectively process information across multiple variables. By dividing the input matrix $x \in \mathbb{R}^{d \times L_1}$ into subsequences or 'patches,' PatchTST captures temporal dynamics with precision. These patches undergo processing through transformer blocks, adept at modeling dependencies along the axes of time and feature dimensionality. Key to PatchTST's architecture are its embedding layer, which elevates the dimensionality of input patches for subsequent processing, and its transformer encoder layers, featuring multi-head self-attention mechanisms and position-wise feed-forward networks. These components enable the modeling of intricate temporal relationships, culminating in an output linearly projected to dimensions $x \in \mathbb{R}^{d \times L_2}$. In our model, we set $L_1 = 48$ and $L_2 = 24$ to predict the next day's patient status using data from the prior 48 hours, employing mean imputation to address irregularities in time-series data.

Training PatchTST necessitates selecting an appropriate loss function, optimizer, and constructing a training regimen. We utilize the Mean Squared Error (MSE) loss for its aptitude in regression tasks, specifically in gauging the accuracy of time-series forecasts. Adam optimizers were chosen for their efficiency with sparse gradients and adaptive learning rates, tested across various initial learning rates ($5e-3, 1e-3, 1e-4, 5e-4$). Additionally, the implementation of a OneCycle learning rate scheduler alongside three other scheduling functions further refines our training process.

For LSTM configurations, we opted for a three-layer setup with 512 hidden dimensions, adhering to standard configurations for both PatchTST and PatchTSMixer to ensure consistency in model evaluation. The effectiveness of these models was

determined based on MSE loss performance on a validation set, with comparative results detailed in Table 1.

C. Detailed Rules for Necessary Lab Test Orders

In our study, we focus on ten blood tests frequently ordered in clinical settings: Complete Blood Count (CBC), Electrolytes, Calcium Profile, INR, Liver Profile, Lactate, Arterial Blood Gas (ABG), Creatinine, Troponin, and Creatinine Kinase (CK). We derived a set of lower bound rules for these test orders after extensive discussions with medical experts, which are encapsulated in the rule set $\mathcal{CR}$ used in Algorithm 1:

- If patient receives blood transfusion, then order CBC, Electrolytes, and INR.
- If patient Urine Output of the last 24 hours is less than 1 liter or greater than 4 liters, order Electrolytes and Creatinine.
- If patient had 25% increasing dose of Vasopressors (or receiving new Vasopressor), order CBC, Liver Profile, Troponin, Lactate.
- If patient had dialysis or will have dialysis, order Calcium Profile.
- If patient has a new fever, order CBC and Liver Profile.
- If the patient Minute Ventilation is increased or decreased by 25%, order ABG.
- If the patient Airway Pressure has 25% increase, then order ABG.
- If the patient had Antibotics treatment, order CBC.
- If the patient had Antiarrhythmics treatment, order Calcium Profile and Electrolytes.
- If the patient had Anticoagulants treatment, order INR.
- If the patient had Propofol treatment, order CK.
- If the patient is on ICP Monitor, order Electrolytes.
- If the patient White Blood Cell (WBC) is less than 1 or greater than 12, order CBC and Liver Profile.
- If the patient White Blood Cell (WBC) has 5 unit of change in the past 24 hours, order CBC.
- If Creatinine value greater than 150 or has 50 increase in the past 24-48 hours, order ABG, Electrolytes, and Calcium Profile.
- If the patient Creatinine Kinase value greater than 5000, order CK.
- If the patient PEEP value has increase more than 2, order ABG.
- If the patient PH is less than 7.3, order Lactate and Creatinine.
- If the patient Hemoglobin value is less than 7, order CBC and INR.
- If the patient Hemoglobin value has decreased more than 2 unit in the past 24 hours, order CBC.
- If patient Platelets is less than 30 or greater than 600000, order CBC.
- If patient Platelets value has more than 30% decrease in the past 48 hours, then order CBC.
- If patient had K replacement in the past 12 hours, order Electrolytes.
- If patient had Ca replacement in the past 12 hours, order Electrolytes.
- If patient had P replacement in the past 12 hours, order Electrolytes.
- If patient had Mg replacement in the past 12 hours, order Electrolytes.

- If patient Sodium (Na) has 6 unit change in the past 24 hours, order Electrolytes.
- If patient Sodium (Na) is greater than 150 or less than 135, order Electrolytes.
- If patient Potassium (K) is greater than 5 or less than 3.5 order Electrolytes.
- If patient Potassium (K) is greater than 4.5, order Creatinine.
- If patient Calcium is greater than 3 or less than 2, order Calcium Profile.
- If patient Phosphate is greater than 0.6 or greater than 1.8, order Calcium Profile.
- If patient Magnesium is greater than 0.8, order Calcium Profile.
- If patient INR is greater than 1.6, order INR.
- If patient Alanine Transaminase (ALT) is greater than 100, order liver profile.
- If patient Bilirubin is greater than 50, oreder liver profile.
- If patient uses Hepatotoxic drug, order liver profile.
- If patient has Arrhythmia, order Troponin and Calcium Profile.
- If patient had Diuretics, order Calcium Profile.

Despite the comprehensive suite of rules applied to our patient ICU stay dataset, it's important to note that these rules were crafted with a high degree of conservatism. This approach is in alignment with clinical practices, ensuring that the rules are seen as necessary and reasonable by healthcare professionals. For each patient stay x in our dataset, we utilize Algorithm 1 to generate a binary vector of length 10, indicating the ordered tests for the following day.

In Figure 3, we present a comparison between the orders generated by our rules and the orders actually placed by physicians. This visualization serves to highlight the extent to which our algorithmically generated orders align with real-world clinical decision-making.

D. Estimate Propensity Score Function with Conditional Normalizing Flow

This section elaborates on the foundational concepts and specific methodologies employed for developing our approximate generalized propensity score (GPS) function, denoted as $\hat{f}(x, t)$.

Normalizing Flows initially emerged to enhance the variational inference in variational autoencoders, as documented in seminal works (Rezende & Mohamed, 2015; Tabak & Vanden-Eijnden, 2010). These models operate by converting a straightforward initial distribution, for instance, a Gaussian, into a complex distribution that closely resembles the distribution of the actual data. This is achieved through a sequence of reversible mappings. We represent the initial distribution by $p_z(z)$, with z being a latent variable, and the distribution of the actual data by $p_x(x)$, where x indicates the observed data. The objective is to discover a function f that facilitates the transformation $x = f(z)$.

At the heart of normalizing flows lies the concept of utilizing a composition of bijective functions $f = f_K \circ f_{K-1} \circ \ldots \circ f_1$, with K indicating the count of these transformations. An affine transformation usually represents the final transformation f_K, while preceding transformations are reversible and nonlinear. Due to the invertibility of each function f_k, the reverse mapping is straightforwardly computed.

The probability density of x in relation to z is calculable through the change of variables theorem. Assuming the base distribution $p_z(z)$ is well-defined and simple (e.g., Gaussian), the density of x can be determined as follows:

$$p_x(x) = p_z(z) \left| \det \left(\frac{\partial f}{\partial z} \right) \right|^{-1}$$

Here, $\left| \det \left(\frac{\partial f}{\partial z} \right) \right|$ signifies the determinant of the Jacobian matrix of transformation f relative to z, illustrating the alterations in the latent space density to align with x's density.

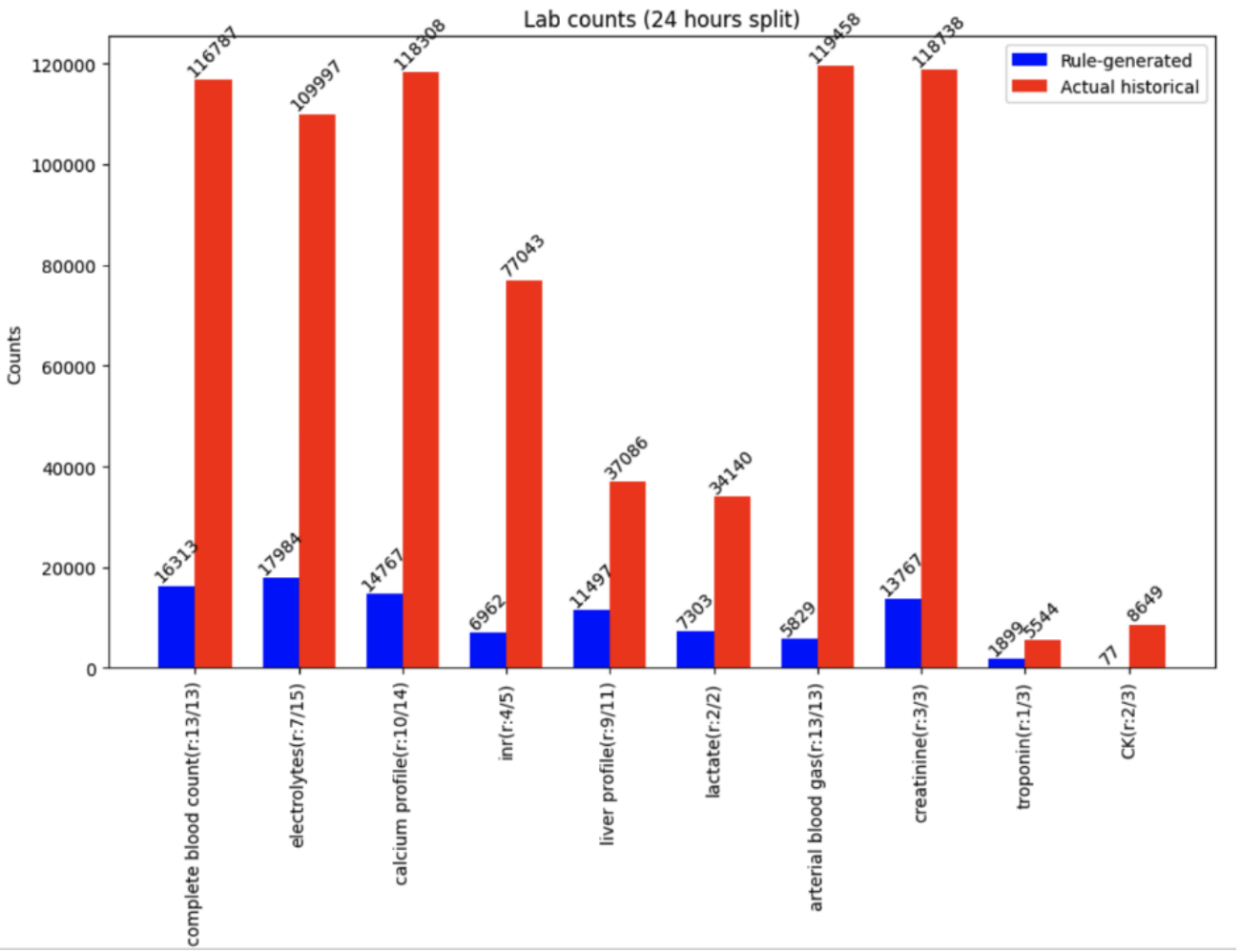

Figure 3. Bar plot that shows the distribution of guideline-generated tests and observed tests

For inference tasks like density estimation or sampling, it is crucial to compute the log-likelihood of the observed data x. Considering a dataset $\mathcal{D} = x_1, \ldots, x_n$, the log-likelihood sums up the log-densities for each data point:

$$\log p(\mathcal{D}) = \sum_{i=1}^{n} \log p_x(x_i).$$

In our implementation, we adopt conditional normalizing flows (CNFs) (Trippe & Turner, 2018; Winkler et al., 2019) following (Schweisthal et al., 2023) for the GPS estimation. CNFs adapt the concept of normalizing flows to conditionally model densities $p(y \mid x)$ by applying an invertible transformation to a base density $p(z)$, with transformation parameters $\gamma(x)$ reliant on the input x.

Utilizing neural spline flows (Durkan et al., 2019) in conjunction with masked auto-regressive networks (Dolatabadi et al., 2020), our methodology enables modeling the conditional distribution of data variables based on a conditioning variable, sequentially generating each variable while considering previously generated variables. This sequential generation underpins efficient computation and sampling from the conditional distribution. CNFs stand out due to their universal approximation capabilities, ensuring accurate density function modeling for complex scenarios, alongside benefits of proper normalization, parametric nature facilitating constant inference time post-training (Melnychuk et al., 2023).

For the GPS modeling $\hat{f}(t, x)$, we integrate neural spline flows with masked auto-regressive networks, setting a flow length of 3, adopting quadratic splines across equally spaced bins. The autoregressive model, a Multilayer Perceptron (MLP) with three hidden layers and 50 neurons each, incorporates noise regularization using noise from $N(0, 0.1)$. The training

Table 5. Testset performance for baseline and our learned policies (with std).

POLICY	$\Delta X \uparrow$	COST$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
RANDOM$_{0.5}$	0.23 $\pm$0.01	0.51$\pm$0.01	4.3$\pm$0.01	3.2$\pm$0.01	1.1$\pm$0.01
RANDOM$_{0.75}$	0.34 $\pm$0.01	0.75$\pm$0.01	3.64$\pm$0.01	1.6$\pm$0.01	2.04$\pm$0.01
LOWERBOUND	0.37	0.62	0	0	0
UPPPERBOUND	0.44	0.82	0	0	0
PHYSICIAN	0.41	0.67	1.24	1.24	0
OURS(W/O GPS)	0.44 $\pm$0.01	0.8 $\pm$0.003	1.06$\pm$0.001	0.34	0.72
OURS(W GPS)	0.42 $\pm$0.005	0.66$\pm$0.005	1.16$\pm$0.001	0.67	0.49

Table 6. Testset performance of prior work and ours policies (with std).

POLICY	$\Delta X \uparrow$	COST$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	INFO GAIN$\uparrow$
PHYSICIAN	0.41	0.67	1.24	0.99
RL (LOW COST)	-	0.62	1.8	1
RL (HIGH COST)	-	0.8	2.3	1.4
OURS(W/O GPS)	0.44$\pm$0.01	0.8 $\pm$0.003	1.06$\pm$0.001	(1.2)
OURS(W GPS)	0.42$\pm$0.005	0.66$\pm$0.005	1.16 $\pm$0.001	(0.98)

of CNFs minimizes the negative log-likelihood (NLL) loss, employing the Adam optimizer with a batch size of 512 over up to 300 epochs, incorporating early stopping based on NLL loss on a validation dataset. Learning rate tuning spans $\{0.0001, 0.0005, 0.001, 0.005, 0.01\}$, with model evaluation mirroring early stopping criteria. For input handling, x, a two-dimensional time-series matrix, is flattened into a vector for processing, ensuring accurate covariate

E. Hyperparameters for Policy Training

For our policy network π_θ, we opt for a PatchTSMixer architecture. We determine the reliability threshold $\bar{\varepsilon}$ as the 5%-quantile of the estimated GPS $\hat{f}(t, x)$ from the training set, unless specified differently. For the optimization of parameters θ and λ, we employ Adam optimizers, considering batch sizes of $\{512, 1024, 2048, 4096\}$. The network is trained leveraging the gradient descent-ascent optimization objective outlined in Eq. (10), targeting a maximum of 50 epochs. Early stopping is implemented based on a patience of 7 epochs for the validation loss, as determined by Algorithm 2 on the factual validation dataset.

The learning rate for updating λ is set to $\eta_\lambda = 0.01$. A random search across 10 configurations is conducted to fine-tune the learning rate for updating the policy network's parameters, η_θ, within the set $\{5e-3, 1e-3, 5e-4, 1e-4\}$, as well as to initialize the Lagrangian multipliers λ_i within the range [1, 5, 10]. Additionally, we explore different values for the outcome function terms, specifically $\beta_1 = \{0, 1, 10, 100\}$ and $\beta_2 = \{0, 1, 10, 100\}$. The performance evaluation during the hyperparameter tuning phase adheres to the same criterion used for early stopping. Subsequent to the hyperparameter determination, we conduct $k = 5$ experimental runs to identify the optimal policy setting.

F. Other Experiments for Policy Evaluation (New)

F.1. Time-series setting

F.2. Original v.s. test guaranteed setting

F.3. RL v.s. Ours in detail

G. Other Experiments for Policy Evaluation

In this section, we delve into additional experimental outcomes, emphasizing the comparative analysis of our formulated policies against standard baselines and reinforcement learning (RL) strategies. Detailed outcomes are encapsulated in Table 5 and Table 6, showcasing the efficacy of our policies relative to conventional approaches.

Our observations reveal that the absence of a generalized propensity score (GPS) does not deter the outcome function's capacity to steer the policy towards achieving a minimized loss, albeit with a tendency to favor policies associated with elevated costs.

For the comparison with RL methodologies, we draw upon the framework of Chang et al. (2019), who employed the final LSTM hidden layer as the state representation x. Aligning our data temporally, we adopt this LSTM layer as our state x, assessing our laboratory test orders as actions at each temporal step. This approach facilitates the evaluation of our policy's effectiveness through the cumulative gain in information, gauged by the discrepancy in probabilities as per their off-policy model.

The off-policy evaluation metric, predicated on the regression of state-action pairs against the differential in mortality classifier probabilities, aims to minimize the informational exposure to users at testing phases. This raises an intriguing query: Does the ordering of lab tests directly correlate with patient mortality? Such an assumption may inadvertently suggest a disparity in clinical treatment across patients.

Exploring Policy Learning without Predicted Future Insights

In pursuit of enhanced explainability and adherence to the logical progression of lab test ordering, we instituted a model for forecasting future patient states, enabling our policy to incorporate anticipated future patient conditions. This not only augments explainability but also highlights a significant reduction in the outcome—specifically, a 20-25% decline in the utility value of the lab usefulness function $g(t, x)$ and in the bound loss, particularly when the forecasted future is excluded from policy training.

Incorporating Real-world Lab Test Costs

We adjusted our model to reflect actual lab test costs as documented in literature, with values delineated as $[12, 5, 12.36, 18, 9.1, 10, 18.62, 1.5, 18, 1.5]$ in USD. The normalization of α_j within our outcome function's cost term leverages this cost array. Policies formulated with this real-world cost paradigm have demonstrated an ability to curtail overall expenses by 5-8% on average. Given an average test cost of \$10.8, a daily ordering volume of 1000 tests could translate into savings of \$500-900, significantly alleviating hospital financial strains and reducing bio-hazardous waste.

Ablation Study on Outcome Function Components

A distinctive aspect of our proposed outcome function $g(t, x)$ is its composition, which encapsulates three key dimensions reflective of optimal lab test ordering practices. Our ablation study on these components reveals their substantial influence on policy formulation.

Eliminating the bound component results in extreme policy behaviors: either an all-inclusive ordering approach to maximize ΔX or a total abstention to minimize costs. Sole reliance on the ΔX component propels the policy towards maximal ordering, culminating in a peak ΔX of 4.52 and a bound loss $L_b^{test} = 2.6$. Conversely, prioritizing cost reduction or assigning significant weight to β_2 leads to a policy of non-ordering, characterized by a zero L_{up} and a maximum $L_{low} = 6.4$.

Thus, the bound term $\mathcal{L}_b$ emerges as pivotal within the outcome function, guiding the policy towards higher cost strategies yet maintaining a threshold (akin to outcomes observed with an Upper Bound policy). Our exploration into the weighting of these terms suggests that a balanced approach yields favorable policy outcomes, though our analysis was confined to integer weight adjustments. Future investigations might benefit from a comprehensive hyperparameter optimization across the β coefficients.