Proceedings of Machine Learning Research 287:1–32, 2025 Conference on Health, Inference, and Learning (CHIL) 2025

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

Zongliang Ji[1,3] JERRYJI@CS.TORONTO.EDU
Andre Amaral[1,2] ANDRECARLOS.AMARAL@SUNNYBROOK.CA
Anna Goldenberg[1,3] ANNA.GOLDENBERG@UTORONTO.CA
Rahul G. Krishnan[1,3] RAHULGK@CS.TORONTO.EDU
[1] *University of Toronto, Canada*
[2] *Sunnybrook Health Sciences Centre, Canada*
[3] *Vector Institute, Canada*

Abstract

Ordering a minimal subset of lab tests for patients in the intensive care unit (ICU) can be challenging. Care teams must balance between ensuring the availability of the right information and reducing the clinical burden and costs associated with each lab test order. Most in-patient settings experience frequent over-ordering of lab tests, but are now aiming to reduce this burden on both hospital resources and the environment. This paper develops a novel method that combines off-policy learning with privileged information to identify the optimal set of ICU lab tests to order. Our approach, EXplainable Off-policy learning with Side Information for ICU blood Test Orders (ExOSITO) creates an interpretable assistive tool for clinicians to order lab tests by considering both the observed and predicted future status of each patient. We pose this problem as a causal bandit trained using offline data and a reward function derived from clinically-approved rules; we introduce a novel learning framework that integrates clinical knowledge with observational data to bridge the gap between the optimal and logging policies. The learned policy function provides interpretable clinical information and reduces costs without omitting any vital lab orders, outperforming both a physician's policy and prior approaches to this practical problem.

Data and Code Availability This paper uses the MIMIC-IV (Johnson et al., 2023) and HiRID (Hyland et al., 2020) datasets, both available on the PhysioNet repository. Code to reproduce the experimental results for ExOSITO is available at this repository[1].

Institutional Review Board (IRB) This study does not require IRB.

1. Introduction

Ordering lab tests is a critical yet challenging task for clinicians, essential for assessing a patient's status and determining the appropriate treatment plan (Zimmerman et al., 1997; Kumwilaisak et al., 2008; Hjortsø et al., 2023). However, studies have shown that clinicians and hospitals often over-order lab tests (Feldman, 2009; Badrick, 2013). This occurs for several reasons: a sense of responsibility, the mental ease of ruling out conditions, and pressures within medical hierarchies (Griffith et al., 1997; Van Walraven and Naylor, 1998; Korenstein et al., 2018; Zhi et al., 2013; Sedrak et al., 2016). Over-ordering generates significant costs for patients, hospitals, and the environment while causing discomfort from unnecessary blood draws, and it may even worsen patient outcomes (Berenholtz et al., 2004; Salisbury et al., 2011). Moreover, redundant tests do not necessarily improve diagnostic accuracy (Iosfina et al., 2013; Pageler et al., 2013). In this work, we propose a novel, explainable tool to assist clinicians in ordering the optimal set of lab tests for intensive care units (ICU) patients.

There are two categories of prior approaches; both face significant challenges that prevent real-world clinical deployment. The first category of work leverages **rule-based interventions**, such as capping the number of tests or reducing tests frequency (Dewan et al., 2016, 2017; Kotecha et al., 2017). These strategies lack adaptability as they do not account for indi-

[1] https://github.com/Jerryji007/ExOSITO-CHIL2025

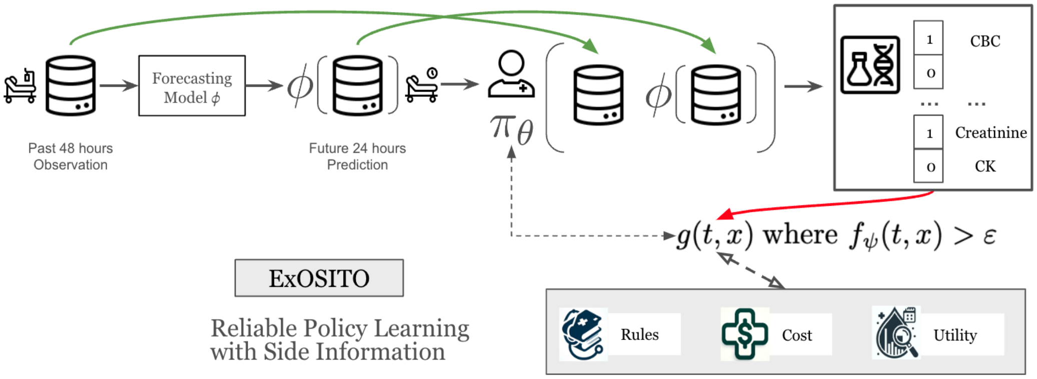
Forecasting Model ϕ
Past 48 hours Observation
ϕ
Future 24 hours Prediction
π_θ
ϕ
1
CBC
0
...
...
1
Creatinine
0
CK
$g(t, x)$ where $f_\psi(t, x) > \varepsilon$
ExOSITO
Reliable Policy Learning with Side Information
Rules
Cost
Utility

Figure 1: Overview of Proposed Method. Top: Development of an ICU patient status forecasting model ϕ for future predictions. Both observed past and predicted future of patient are inputs of policy π_θ (green arrow), which determines the next-day lab test order for the patient. Bottom: π_θ is learned by maximizing the reward function g, which evaluates each test order (red arrow) and incorporates three main components (dashed arrow) along with f_ψ, the learned global propensity score (GPS) function, to ensure non-trivial support.

vidual patient characteristics or changing clinical conditions over time. This may lead to the omission of necessary tests and compromise care (Kobewka et al., 2015). Another line of work (Cheng et al., 2018; Chang et al., 2019; Ji et al., 2024) uses offline, off-policy **reinforcement learning (RL)**. While these methods offer more adaptive decision-making by optimizing test ordering based on historical data, they rely on objectives that do not align with decision-making dynamics in clinical settings. First, they aim to replace clinical oversight – a non-starter in practical deployments for high-risk medical settings such as ICUs since the legal and medical imperative for decision lies firmly with clinicians (Lu et al., 2020). Our strategy instead is to focus on creating assistive tools for clinicians within the ICU. This requires reframing the learning problem from developing an RL agent meant to replace the clinician to a bandit formulation that offers real-time assistive support to the attending physician. Next, prior methods rely solely on historical data for decision making, making them susceptible to learn policies that are not robust to changing standards of care; e.g. some hospitals may have previously allowed more frequent test orders than current practices permit, leading to significant discrepancies between the offline policy and the intended real-world deployment scenario. This results in low-overlap between the offline and online policy. Finally, a significant concern during deployment is interpretability which is crucial for clinicians to build trust when using the tool during silent deployment. These challenges motivate our work.

Contributions: Our strategy for improving the robustness of learned policies from offline data is to leverage *physiologically grounded* privileged information (information that is available at training time, but not at test time) in the form of clinical rules. This enables us to mitigate the difference between the optimal and logging policy. To improve explainability, we provide practitioners a proxy of the learner's assessment of the *current* patient state through the use of forecasts that visualize the expected future evolution for the patient.

By combining these ideas we propose EXplainable Off-policy learning with Side Information for ICU blood Test Orders (ExOSITO); an interpretable assistive tool for clinicians to order lab tests using both the observed and predicted future status of a patient. To the best of our knowledge, this is the first work to use privileged information to create an explainable, clinician-facing tool for lab test ordering, leveraging verified clinical rules as side information within an off-policy bandit framework. We study the framework on real-world medical datasets and showcase how the hybrid framework outperforms both a physician's policy and prior approaches bringing us significantly closer to a deployable clinical decision support tool.

2. Related Work

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 do not generalize to other contexts. 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. Soleimani et al. (2017) used regression methods to model the novel information each test provides. These approaches focus on specific diseases and overlook a large number of clinical factors, such as predictive information from vital signs and the causal links between test orders and their utility to clinicians. Such omissions make the model less adaptable to variations in patient conditions and unsuitable for deployment in an ICU.

Off-policy learning, particularly via RL, has been extensively explored for ICU patient treatment plans (Komorowski et al., 2018; 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, such methods typically aim to maximize patient outcomes as the reward (See Figure 3). The challenge with this formulation is that lab test orders should be viewed as actions that provide information to clinicians rather than as having a direct impact on patient outcomes. Ignoring the clinician as the key decision maker in the loop breaks the assumption underlying the Markov decision process. This motivates us to adopt a conceptually simpler bandit formulation aimed at providing instantaneous assistance to clinicians in deciding which lab tests to order. Cheng et al. (2018) focused on sepsis-related lab tests using only four lab features limiting its generalizability to broader ICU settings. Chang et al. (2019) applied deep Q-learning to frequently measured features, such as vital signs but their work omits ordering for several blood test features. Ji et al. (2024) built on Chang et al. (2019)'s framework by applying updated learning methods but found no significant differences across learned policies due to limitations in the logging policy and the use of mortality as the sole reward.

Off-Policy Learning with Constraints and Side Information: Off-policy learning refers to learning a strategy for making decisions, called a policy, using data that was collected under a different task and/or goals. (Lange et al., 2012; Sutton and Barto, 2018; Levine et al., 2020). Off-policy learning, without further interaction with the environment, is challenging, as the policy the data exhibits may not be optimal. This issue is evident in lab test ordering, where clinicians err on the side of ordering redundant tests. Previous studies have shown that learned policies can 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 implicit regularization in the form of generative models to estimate the behavioral policy (Zhou et al., 2021; Ghasemipour et al., 2021) or by incorporating regularization penalties to constrain the divergence between the learned and behavior policies (Jaques et al., 2019; Kumar et al., 2019; Wu et al., 2019, 2022; Mao et al., 2023). Peripherally related is the field of Bayesian Optimization (Yang et al., 2022; Wilson et al., 2014; Letham and Bakshy, 2019; Müller et al., 2021; Cheng et al., 2018) which blends active and policy learning with the goal of maximizing an expected reward. However, Bayesian Optimization requires access to a simulator and precise simulators are rarely available in healthcare. Le et al. (2019) studies learning such policies with multiple constraints, and Schweisthal et al. (2023) leverages ideas from causal inference; specifically, propensity score estimation to guide the design of policies to ensure non-trivial overlap over covariate-treatment space. Our work builds on these ideas.

Side information has been used in off-policy evaluation (Felicioni et al., 2022) as well inverse reinforcement learning (Wen et al., 2017) but not in a deployment focused application. In this work, we showcase the impact that even simple rules summarized by clinicians can, as side information, guide policy learning.

3. Background

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.[2] $x_i \in X \subseteq \mathbb{R}^{d\times L}$ is the context, covariate, or a patient ICU stay represented by an irregular time-series matrix, where d is the number of features and L is the length of stay. $t_i \in T \subseteq \{0,1\}^K$ is the action, or the lab test order represented by a K-dimensional binary vector.

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

$y_i \in Y \subseteq \mathbb{R}$ is the reward, outcome, or the utility of the lab test order represented by a real number. We highlight that T *does not* refer to the drugs or clinical interventions that a patient may be prescribed within the ICU – one of the assumptions that we make in this work is that the clinical covariates, X, capture the effect of any medication/clinical intervention and consequently sufficient evidence to decide upon which lab test to order.

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). 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:

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

As it is understood that clinicians' planning and actions vary, we assume that the logging policy represents a sub-optimal policy within the set Π. Since our problem operates under an off-policy setting, timely feedback on $Y(\cdot)$, which represents the effects of lab test orders on clinicians, is not available. In our method, we use the *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. Prior work models g as patient mortality estimation, which, while easily obtainable, is not informative for our problem (Chang et al., 2019; Schweisthal et al., 2023). Our work creates a novel variant of $g(t, x)$, a closed-formed differentiable function that we refer to as the *lab order utility function.*

Assumptions: We assume that the missingness pattern that the longitudinal data exhibit is missing at random. For the outcome estimation to be identifiable, we adhere to three standard assumptions in causal inference (Rubin, 1974): (1) *Consistency* ($Y = Y(T)$ [3]), asserting that observed outcomes align with potential outcomes under the observed treatment. (2) *Ignorability* ($Y(t) \perp\!\!\!\perp T \mid X,\ \forall t \in T$), confirming the absence of hidden confounders [4]. (3) *Overlap* ($f(t, x) > \varepsilon,\ \forall x \in X, t \in T$, 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 $X = x$. 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.

Limited Overlap: 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 utility 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. Our work tackles 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 $X = x$. Following Schweisthal et al. (2023), we opt for *conditional normalizing flows* (CNFs) (Trippe and Turner, 2018; Winkler et al., 2019) to form a parametric estimate of the GPS function. CNFs, built on the foundation of normalizing flows (Tabak and Vanden-Eijnden, 2010; Rezende and Mohamed, 2015), are parametric generative models capable of modeling conditional densities $p(y|x)$. 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, $\mathcal{L}_{\text{nll}} = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}(t, x)$[5]. Our method for policy learning uses the learned GPS function $\hat{f}(t, x)$, steered away from areas of high uncertainty, improving the reliability of the resulting policy.

[3] Abusing notation here, function $Y(\cdot)$ means the potential outcome and Y means actual outcome.

[4] We assume that our set of covariates, which includes treatments (for example drugs and procedures), vital signals, and test results, is complete. A detailed discussion on the validity of this assumption can be found in Appendix A.1.

[5] Additional details on the CNF training process are provided in the Appendix E.

4. Explainable Off-policy learning with Side Information for Test Orders (ExOSITO)

Clinicians are trained through decades of formal education and experience, planning their actions based on established practices. However, in real clinical settings, especially under the pressure of high patient loads, clinicians may resort to ordering as many tests as a precautionary measure, often driven by a heightened sense of responsibility. By incorporating basic rules and knowledge accumulated from years of practice, our approach offers a second opinion, encouraging clinicians to reconsider their initial test orders. This integration of foundational rules with clinical practice helps streamline decision-making and ensures more targeted and efficient patient care.

Summary We introduce an explainable approach to creating a learned policy for ordering laboratory tests in the ICU (refer to Figure 1). If every clinician had the bandwidth to stop and think about what lab tests to order, then one of the decision thresholds they might use is whether or not they think the patient is on a trajectory to worsening or recovery. To leverage this insight we build a forecasting model to predict a patient's future physiological status using their laboratory biomarkers as a proxy for the same. In conjunction with a patient's immediate history, these predictions then form the inputs for our policy function. Next, we establish bounds for each observed lab test order based on clinically validated guidelines. We identify the minimum sets of orders to make by applying clinician-curated rules, and determined the maximum sets of orders by combining the observed and rule-generated orders. These rules are used in the design of a potential outcome function to evaluate the effectiveness of each potential lab test order. They help us mitigate the disparity between physician policy (i.e. logging policy) and our learned policy π. By using these rules as side information in the reward function, we can regularize the off-policy learning algorithm to ensure that it learns a safe and reliable policy for lab test ordering in ICU settings.

4.1. Forecasting for explainability

Previous studies (Cheng et al., 2018; Chang et al., 2019; Ji et al., 2024) represent patient covariates X as an imputed, irregular time-series matrix of the patient's history. Our interactions with medical experts revealed that when clinicians order lab tests with forethought, they attempt to infer insights into a patient's future condition; i.e. they focus not only on current patient states but also on forecasting future lab results when making their prognoses about the patient. This insight prompted the inclusion of both observed and predicted patient states in our representation of patient covariates. To predict patient future status, we first build our forecasting model, ϕ, which is based on PatchTST (Nie et al., 2022). ϕ takes the observed patient status X_{prev} as input. This input comprises a broad spectrum of features, including vital signs, treatments, and relevant lab test results based on recommendations by ICU clinicians. The model is trained to minimize the mean squared error, $\mathcal{L}_{mse} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_{post} - x^*_{post})^2$, between the observed future status X^*_{post} and the predicted future status $\hat{X}_{\text{post}} = \phi(X_{\text{prev}})$[6]. By constructing the context X with patient past and predicted future status, we can learn policies that are explainable to clinicians as each policy output correspond with exact values of patient status. This setup can provide clinicians a better chance to evaluate policy actions during future deployments for 'online' approvals.

Algorithm 1 Obtain bounds for observed test orders

```
Input: Patient stay $x$, set of clinical rules $\mathcal{CR} = \{r^1, \dots, r^M\}$
Output: Minimal $t^{min}$ and maximal $t^{max}$ of lab test order of stay $x$
Determine observed test order $t^* \in \{0,1\}^K$ from $x$
$t^{max}, t^{min} \leftarrow \mathbf{0} \in \{0,1\}^K$, where $K$ is the number of lab tests
foreach $r^m \in \mathcal{CR}$ do
    if $x$ satisfies $r^m$ then
        Find the indices $\mathcal{I} \subset \{1, \dots, K\}$ of lab tests corresponding to rule $r^m$
        $t_i^{min} \leftarrow 1$ for $i \in \mathcal{I}$
    end
    $t^{max} \leftarrow \{t_j^* \vee t_j^{min}\}_{j=1}^K$
end
```

4.2. Minimal and maximal expectations of lab tests

Lab test ordering in ICU is a well-established practice, underpinned by decades of clinical experience, which has yielded straightforward guidelines for test ordering (Kumwilaisak et al., 2008; Cismondi et al., 2013; Vidyarthi et al., 2015; Bindraban et al., 2018). 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. Despite this, the choices made regarding which labs

[6] Details about the construction of our patient status forecasting model are provided in Appendix B.

to order in the physician policy may not represent the optimal set. But why learn a policy when such guidelines exist? The answer is twofold: (1) These guidelines are basic, derived from historical knowledge, and tend to be exceedingly conservative, being triggered only when patients meet certain extreme criteria. (2) These guidelines apply to both past and future patient states. For example, if a clinician knows with certainty that a patient's hemoglobin will drop below a threshold tomorrow, a CBC should be ordered preemptively. However, clinicians cannot accurately predict future patient conditions in practice. Despite this, such guidelines are uniquely suited to be incorporated as priors in policy learning systems, as they are inherently conservative and can be combined with forecasting tools.

To effectively incorporate clinical guidelines into our policy learning framework, we derive stay-specific bounds that constrain the possible set of lab tests to be ordered. For each patient stay x, we can determine a conservative bound $t^{min} \in \{0,1\}^K$ 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 (Wang et al., 2017; Sachdeva et al., 2020), we establish a permissive bound of each test order as $t^{max} = \{t^*_j \vee t^{min}_j\}_{j=1}^K \in \{0,1\}^K$, which is a combination of the guideline-derived order t^{min} 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. For a rule stating, 'If Hemoglobin is less than 7, order CBC.' For a given patient stay x, if a Hemoglobin measurement is less than 7, then $t^{min}_{CBC} = 1$. Due the conservative nature of the clinical guidelines, the guideline-generated orders constitute about 30% of the observed orders. Further details on the rule generation process and supporting literature for these clinical guidelines can be found in Appendix C.

4.3. Potential Outcome Function for Lab Test Order Utility

Since clinicians plans (rules) and actions (observations) differ, the treatments (labs) in the 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 assesses 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 three characteristics of an effective lab test order:

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

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

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

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

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: $\Delta_{range}(t, x) = \sum_{j=1}^{K} \mathbb{1}(t_j > 0.5) \cdot \max(\delta_{max}, \delta_{min})$, where $\delta_{max} = |\max(x^{lab,t_j}_{prev}) - \max(x^{lab,t_j}_{post})|$ and $\delta_{min} = |\min(x^{lab,t_j}_{prev}) - \min(x^{lab,t_j}_{post})|$ are the absolute differences between the maximum and minimum values of predicted and observed values for test t_j. Our decision to focus on average and extreme value differences rather than variance followed clinical consultations, favoring metrics that assess whether features meet or exceed clinically meaningful thresholds, aligning closely with practical needs in real-world settings. The clinical and literature support of our design on ΔX is further illustrated in Appendix D.

b] Safe, yet useful: Lab test orders are encouraged to be safe in terms of meeting conservative requires laid out by the rules, while should conform to defined in Sec 4.2. The deviation of each test order from the bounds t_j^{min} and t_j^{max} is quantified by $\mathcal{L}_b$:

$$\mathcal{L}_b(t, x) = \sum_{j=1}^{K} \mathbb{1}(t_j^{max} = t_j^{min}) \cdot |t_j^{min} - t_j|. \tag{3}$$

The indicator function ensures the policy includes all necessary labs suggested by rules while avoiding redundant ones. A lower value of $\mathcal{L}_b$ indicates that the test order t is better aligned with the criteria of meeting the minimal and maximal expectations for orders. Here we treat $t \in [0, 1]^K$ as the predicted probability for each test for our policy.

c] Cost effective: 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; Ji et al., 2024) that assume uniform cost across tests, we consider the relative clinical costs of each test. The cost function is defined as:

$$C(t) = \sum_{j=1}^{K} \alpha_j \cdot \mathbb{1}(t_j > 0.5), \sum \alpha_j = 1, \tag{4}$$

where α_j represents the cost associated with each lab test. Synthesizing these desirable qualities, we define the lab test order utility function (conditional outcome function) as:

$$g(t, x) = \Delta X(t, x) - \beta_1 \mathcal{L}_b(t, x) - \beta_2 C(t), \tag{5}$$

with β_1 and β_2 as regularization hyperparameters. Next, we show how we use the GPS function, $g(t, x)$ to estimate the policy value $\hat{V}(\pi)$ during policy learning.

4.4. Overlap-Guaranteed Policy Learning

We present a time-aware, overlap-guaranteed off-policy learning algorithm designed for ICU lab test ordering, prioritizing explainability and reliability. This algorithm seeks to optimize policy π^{rel} to maximize the estimated policy value $\hat{V}(\pi)$ while ensuring substantial data support in the covariate-treatment domain to guarantee the overlap condition.

The policy class $\Pi^{\text{r}} = \{\pi \in \Pi \ \mid f(\pi(x), x) > \bar{\varepsilon}\}, \forall x \in X\}$ is defined to include only those policies that maintain a minimum overlap, specified by a reliability threshold $\bar{\varepsilon}$. We then reformulate our objective from Eq. (1) as: $\pi^{\text{rel}} \in \arg\max_{\pi \in \Pi^{\text{r}}} \hat{V}(\pi)$. Given the finite nature of our observational data, we formulate the optimization as:

$$\max_{\pi} \quad \frac{1}{n} \sum_{i=1}^{n} g(\pi(x_i), x_i) \quad \text{s.t.} \quad \hat{f}(\pi(x_i), x_i) \geq \bar{\varepsilon} \tag{6}$$

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's next 24 hours based on the
  past 48 hours status
// Step 2:  Find lab order bounds using Algorithm 1 and
   define lab test utility function
Prepare $t^{max}$, $t^{min}$ 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 $m = 1$ to $M$ do
    for each epoch do
        for each batch do
            // Predict next 24 hours ICU stay with
               forecasting model
            $x_{post} \leftarrow \phi(x_{prev})$
            $x_i \leftarrow [x_{prev}, x_{post}]$
            $\mathcal{L}_\pi \leftarrow -\frac{1}{n} \sum_{i=1}^{n} \cdot$
              $\left\{ g\left(\pi_\theta^{(m)}(x_i), x_i\right) - \lambda_i \left[\hat{f}\left(\pi_\theta^{(m)}(x_i), x_i\right) - \bar{\varepsilon}\right] \right\}$
            $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_\pi$
            $\lambda \leftarrow \lambda + \eta_\lambda \nabla_\lambda \mathcal{L}_\pi$
        end
    end
end
// Select best learned policy with respect to constrained
   objective on validation set
$\pi_\theta^{\text{rel}} \leftarrow \pi_\theta^{(m^*)}$, with
    $m^* = \arg\max_m \sum_{i=1}^{n} g\left(\pi_\theta^{(m)}(x_i), x_i\right) \cdot \mathbb{1}\left\{\hat{f}\left(\pi_\theta^{(m)}(x_i), x_i\right) > \bar{\varepsilon}\right\}$
```

Here the lab test order utility function g serves as our policy outcome estimator, and the estimated GPS $\hat{f}$ limits the policy search space. To facilitate learning, we employ neural networks parameterized by θ and reformulate the constrained problem in Eq. 6 into an unconstrained Lagrangian form:

$$\min_{\theta} \max_{\lambda_i \geq 0} -\frac{1}{n} \sum_{i=1}^{n} \left\{ g\left(\pi_\theta(x_i), x_i\right) - \lambda_i \left[\hat{f}\left(\pi_\theta(x_i), x_i\right) - \bar{\varepsilon}\right] \right\}, \tag{7}$$

where λ_i are Lagrange multipliers.

This adversarial learning approach, utilizing gradient descent-ascent techniques (Lin et al., 2020), enables the establishment of a robust policy under the defined constraints. The implementation specifics

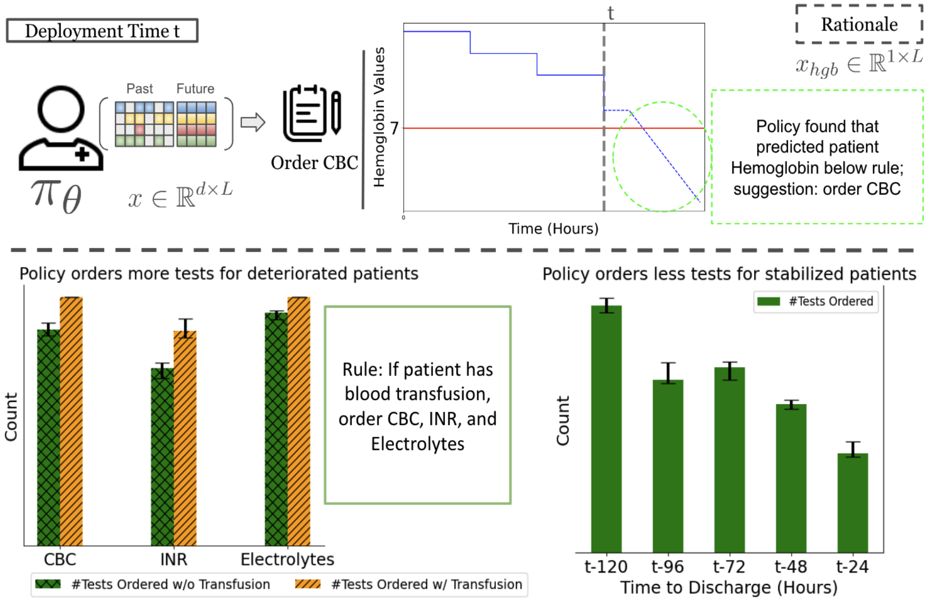

Figure 2: Integration of Medical Knowledge into Explainable Policy Learning for real patient data. Top: At deployment, our policy transparently justifies actions, such as ordering a CBC test, based on predictions like a decrease in future Hemoglobin levels. Bottom: The policy incorporates clinical guidelines for lab test ordering. At test time, the policy increases test orders for deteriorating patients due to changes in transfusion features, resulting in more corresponding tests being ordered (left). Conversely, it reduces test orders for stabilized patients as discharge approaches (right).

and algorithmic details are further elaborated in Algorithm 2.

5. Empirical Evaluation

We conduct comprehensive experiments on two real-world ICU patient datasets to assess our method. **MIMIC and HIRID:** The MIMIC-IV database (Johnson et al., 2023) contains anonymized health records from patients who stayed in the ICU of a U.S. hospital. The HiRID (Hyland et al., 2020) is a publicly available critical care dataset that includes high-resolution data from patients admitted to an ICU in Switzerland. 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 for MIMIC-IV, each characterized by 71 irregular time-series features and $n = 32,216$ patient stays from HiRID, each with 73 features [7] mirroring clinician daily practices. These features encompass lab test results, vital signs, and patient treatments. Further details on the preprocessing of the MIMIC-IV and HiRID are provided in Appendix A.

5.1. Patient Status Forecasting

We train a multivariate time-series forecasting model based on observed ICU stays to obtain predicted future patient statuses. The dataset, denoted as $\mathcal{D} = \{x^i_{prev}, x^{i*}_{post}\}^n_{i=1}$, 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 PatchTST (Nie et al., 2022) for our patient status forecasting model ϕ. The training of ϕ focuses on minimizing the mean squared error between the predicted future status x^i_{post} and observed future [8].

[7] The EHR systems for both datasets differ, resulting in a variation in the number of features considered.

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

Table 1: Test set performance for baseline and our learned policies.

	MIMIC-IV					HIRID				
Policy	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
Random$_{0.5}$	0.23	**0.51**	4.3	3.2	1.1	0.47	**0.49**	4.4	2.9	1.5
Random$_{0.75}$	0.34	0.75	3.64	1.6	2.04	0.58	0.74	3.77	1.66	2.11
LowerBound	0.37	**0.62**	0	0	0	0.62	**0.35**	0	0	0
UpperBound	**0.44**	0.82	0	0	0	**1.13**	0.59	0	0	0
Physician	0.41	0.67	1.24	1.24	0	0.96	0.55	0.98	0.98	0
Ours(w/o GPS)	**0.44**	0.8	**1.06**	**0.34**	0.72	**1.08**	0.57	**0.62**	**0.3**	**0.32**
Ours(w GPS)	**0.42**	**0.66**	**1.16**	**0.67**	**0.49**	**1.01**	**0.52**	**0.89**	**0.5**	0.39

Table 2: Test set performance of prior work and ours policies.

	MIMIC-IV				HIRID			
Policy	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	info gain↑	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	info gain↑
Physician	0.41	0.67	1.24	0.99	0.96	0.55	0.98	1.03
RL(Chang et al., 2019)(low cost)	-	**0.62**	1.8	1	-	**0.51**	1.4	1
RL(Chang et al., 2019)(high cost)	-	0.8	2.3	**1.4**	-	0.74	3.6	**1.19**
Ours(w/o GPS)	**0.44**	0.8	**1.06**	**(1.2)**	**1.08**	0.57	**0.62**	**(1.17)**
Ours(w GPS)	**0.42**	**0.66**	**1.16**	(0.98)	**1.01**	**0.52**	**0.89**	**(1.05)**

5.2. Policy Training and Evaluation

Our dataset was partitioned into training, validation, and test sets at 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 $\pi_\theta^{\text{rel}} : \mathbb{R}^{72\times71} \to [0, 1]^{10}$ that yields the highest outcome on the GPS-constrained validation set. We then evaluate the policy with the best validation set performance on test set with our outcome function $g(t, x)$. Further training specifics, including hyperparameters, are detailed in Appendix E & F.

5.3. Results

Our model exceeds all baseline methods. To date, no studies have directly focused on deployment-guided off-policy learning for ICU lab test ordering. However, as each test order t is a binary vector, we compared our trained policies against random, lower and upper bound policies, and, crucially, the observed clinician policy (Table 1). With increasing orders and costs, the information ΔX 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 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 learning algorithm is able to discover policies with higher ΔX and lower out-of-bound test orders $\mathcal{L}_b$ at minimal cost. The ability to consistently find a policy that reduces costs, provides more information, and adheres more closely to clinical guidelines compared to physician policies in historical data highlights the promise of our method as a potential first clinically deployed clinician-facing tool for ICU lab test ordering.

Our GPS approach for reliable policy learning outperforms non-GPS methods and the Physician policy. Our methodology ensures the discovery of reliable policies based on our causal assumptions and the estimated GPS function. As shown in Table 1, 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.

EXOSITO matches RL approaches in information gain while reducing ordering costs and minimizing out-of-bound orders. Although our approach develops a clinician-facing tool within causal contextual bandits, rather than a patient-facing setup based on RL settings where 'next state' is observed, we find it necessary to compare our results with the work of Chang et al. (2019). 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 2, 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 without given any mortality information. This further underscores that mortality is an unsuitable reward metric for lab test ordering.

Rule integration enables learning transparent, clinically grounded policies. 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 H, underscore the policy's capacity to integrate critical clinical insights, enhancing its applicability and trustworthiness in a healthcare setting.

Integrating forecasting, clinical rules, and cost is crucial for robust and minimal lab ordering policies. Our reward function consists of three components that depend on the performance of our learned patient status forecasting time-series model (ΔX), clinical rules as privileged information ($\mathcal{L}_b$), and the relative clinical cost $C(t)$. We perform a detailed set of ablation studies, including ablating elements of the outcome function, varying the accuracy of the patient forecasting model, changing the conservativeness of the clinical rules, and switching between real-relative cost and uniform cost (see Appendix H.1 to H.6). We found that each element of our outcome function is crucial for policy learning and that our learned policy benefits from including patient status forecasting, clinical rules, and real-relative costs. Moreover, our time-series model performs solidly in predicting patient status, as evaluating our learned policy with perfect predictions yields results that are not significantly different from those using our learned model's predictions. Our ablation study also shows that having rules help the policy generate minimal lab orders and having less conservative rules (i.e. more 1's in the t^{lower}) would revert the policy back to the logging policy which errs on over-ordering.

6. Discussion and Future Work

In this paper, we introduce, ExOSITO, 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. We also demonstrate how to leverage privileged information in the form of domain expertise from clinicians to improve the process of learning causal bandits for ICU lab test ordering. Nonetheless, challenges such as ignorability may persist due to potential unobserved confounders and the presence of noisy or corrupted data. Although the EHR dataset encompasses a comprehensive range of clinically relevant variables, unobserved confounding is inevitable given the complexity of the ICU environment. We plan to incorporate as many clinically verified covariates as possible into our policy learning framework to mitigate these issues.

A critical element of our approach is the accurate forecasting of patient future status. Future improvements might include integrating graph-based time-series models to enhance the accuracy and interpretability of patient status predictions. Alongside structured knowledge, the role of missingness in observed irregular time-series patient data is pivotal for forecasting accuracy, prompting us to assume our data is Missing At Random (MAR) to mitigate its effect.

Off-policy Evaluation (OPE) is not used in our setting for two primary reasons. First, our dataset does not include 'true' outcome measures that quantify the utility of lab test orders, which are essential for mapping state-action pairs to clinical outcomes. Second, the logging policy that generated our data is suboptimal and fails to cover the full range of lab test orders, resulting in insufficient support in action (T) space. Together, these limitations prevent OPE from providing a reliable evaluation of our learned policy.

The trained policies of study does not account for potential distribution shifts. 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.

By reframing the problem as a causal bandit and demonstrating its effectiveness relative to prior methods on real-patient datasets, our work establishes a robust foundation for subsequent later empirical studies. 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. With our forecasting setup, we are able to modify the selected features based on the clinicians need. Data collected from such deployments would be invaluable for refining policy accuracy through ground truth labeling and counterfactual learning.

While our results outperform physician orders, we acknowledge that our formulation relies on strong assumptions—namely temporal independence of test ordering decisions and the absence of unobserved confounders—and we will explore semi-Markov decision processes and augment Equation 3 with additional safety terms to relax these constraints and enhance robustness. We also plan to rigorously validate our informativeness metric and extend our cost modeling to capture dynamic factors such as test urgency and resource constraints, ensuring clinically important tests are appropriately valued. Finally, to assess real-world performance and deployment readiness, we will conduct a silent trial in ICU settings to gather direct clinician feedback, compare against rule-based decision support systems, and systematically analyze potential failure modes of our framework.

References

Edris M Alkozai, Bakhtawar K Mahmoodi, Johan Decruyenaere, Robert J Porte, Annemieke Oude Lansink-Hartgring, Ton Lisman, and Maarten W Nijsten. Systematic comparison of routine laboratory measurements with in-hospital mortality: Iculabome, a large cohort study of critically ill patients. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 56(7):1140–1151, 2018.

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

Sean M Berenholtz, Peter J Pronovost, Pamela A Lipsett, Deborah Hobson, Karen Earsing, Jason E Farley, Shelley Milanovich, Elizabeth Garrett-Mayer, Bradford D Winters, Haya R Rubin, et al. Eliminating catheter-related bloodstream infections in the intensive care unit. *Critical care medicine*, 32(10):2014–2020, 2004.

Renuka S Bindraban, Maarten J Ten Berg, Christiana A Naaktgeboren, Mark HH Kramer, Wouter W Van Solinge, and Prabath WB Nanayakkara. Reducing test utilization in hospital settings: a narrative review. *Annals of laboratory medicine*, 38(5):402–412, 2018.

Tuba Damar Çakırca, Gökhan Çakırca, Ayşe Torun, Ahmet Bindal, Murat Üstünel, and Ahmet Kaya. Comparing the predictive values of procalcitonin/albumin ratio and other inflammatory markers in determining covid-19 severity. *Pakistan Journal of Medical Sciences*, 39(2):450, 2023.

Chun-Hao Chang, Mingjie Mai, and Anna Goldenberg. Dynamic measurement scheduling for event forecasting using deep rl. In *International Conference on Machine Learning*, pages 951–960. PMLR, 2019.

Li-Fang Cheng, Niranjani Prasad, and Barbara E Engelhardt. An optimal policy for patient laboratory tests in intensive care units. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 320–331. World Scientific, 2018.

Federico Cismondi, Leo A Celi, André S Fialho, Susana M Vieira, Shane R Reti, Joao MC Sousa, and Stan N Finkelstein. Reducing unnecessary lab testing in the icu with artificial intelligence. *International journal of medical informatics*, 82(5): 345–358, 2013.

Maya Dewan, Jorge A. Gálvez, Tracey Polsky, Genna Kreher, Blair Kraus, Luis M. Ahumada, John J. McCloskey, and Heather Wolfe. Reducing unnecessary postoperative complete blood count testing in the pediatric intensive care unit. *The Permanente Journal*, 2016.

Maya Dewan, Jorge Galvez, Tracey Polsky, Genna Kreher, Blair Kraus, Luis Ahumada, John McCloskey, and Heather Wolfe. Reducing unnecessary postoperative complete blood count testing in the pediatric intensive care unit. *The Permanente journal*, 21, 2017.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Invertible generative modeling using linear rational splines. In *AISTATS*, 2020.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *NeurIPS*, 2019.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Kevin P Eaton, Kathryn Levy, Christine Soong, Amit K Pahwa, Christopher Petrilli, Justin B Ziemba, Hyung J Cho, Rodrigo Alban, Jaime F Blanck, and Andrew S Parsons. Evidence-based guidelines to eliminate repetitive laboratory testing. *JAMA internal medicine*, 177(12):1833–1839, 2017.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. *arXiv preprint arXiv:2306.09364*, 2023.

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

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

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

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *International Conference on Machine Learning*, pages 3682–3691. PMLR, 2021.

Charles H Griffith, John F Wilson, Nirmala S Desai, and Eugene C Rich. Does pediatric housestaff experience influence tests ordered for infants in the neonatal intensive care unit? *Critical care medicine*, 25(4):704–709, 1997.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 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.

C. J. Hjortsø, M. Møller, A. Perner, and A. C. Brøchner. Routine versus on-demand blood sampling in critically ill patients: A systematic review*, 2023.

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

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

Ioulia Iosfina, Hayley Merkeley, Tara Cessford, Georgia Geller, Neda Amiri, Nazli Baradaran, Monica Norena, Najib Ayas, and Peter M Dodek. Implementation of an on-demand strategy for routine blood testing in icu patients. In *D23. QUALITY IMPROVEMENT IN CRITICAL CARE*, pages A5322–A5322. American Thoracic Society, 2013.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Zongliang Ji, Anna Goldenberg, and Rahul G Krishnan. Measurement scheduling for icu patients with offline reinforcement learning. *arXiv preprint arXiv:2402.07344*, 2024.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

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.

Vijay Kandalam, Cheryl K Lau, Maggie Guo, Irene Ma, and Christopher Naugler. Inappropriate repeat testing of complete blood count (cbc) and electrolyte panels in inpatients from alberta, canada. *Clinical Biochemistry*, 77:32–35, 2020.

Ruth M Kleinpell, J Christopher Farmer, and Stephen M Pastores. Reducing unnecessary testing in the intensive care unit by choosing wisely. *Acute and Critical Care*, 33(1):1, 2018.

Daniel M Kobewka, Paul E Ronksley, Jennifer A McKay, Alan J Forster, and Carl van Walraven. Influence of educational, audit and feedback, system based, and incentive and penalty interventions to reduce laboratory test utilization: a systematic review. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 53(2):157–183, 2015.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Flemming Kondrup, Thomas Jiralerspong, Elaine Lau, Nathan de Lara, Jacob Shkrob, My Duc Tran, Doina Precup, and Sumana Basu. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15696–15702, 2023.

Deborah Korenstein, Solomon Husain, Renee L Gennarelli, Cilian White, James N Masciale, and Benjamin R Roman. Impact of clinical specialty on attitudes regarding overuse of inpatient laboratory testing. *Journal of hospital medicine*, 13(12): 844–847, 2018.

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

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Kanya Kumwilaisak, Alberto Noto, Ulrich H Schmidt, Clare I Beck, Claudia Crimi, Kent Lewandrowski, and Luca M Bigatello. Effect of laboratory testing guidelines on the utilization of tests and order entries in a surgical intensive care unit. *Critical care medicine*, 36(11):2993–2999, 2008.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pages 45–73. Springer, 2012.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

Benjamin Letham and Eytan Bakshy. Bayesian optimization for policy search via online-offline experimentation. *Journal of Machine Learning Research*, 20(145):1–30, 2019.

Olga Levi, Maverick Chan, Thomas Bodley, Smith Orla, Michelle Sholzberg, Shannon Swift, Hina Chaudhry, Jan O Friedrich, and Lisa K Hicks. Reducing repetitive and reflexive diagnostic phlebotomy in an intensive care unit: a quality improvement project. *Blood*, 134:3406, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020.

MingYu Lu, Zachary Shahn, Daby Sow, Finale Doshi-Velez, and H Lehman Li-wei. Is deep reinforcement learning ready for practical applications in healthcare? a sensitivity analysis of duel-ddqn for hemodynamic management in sepsis patients.

In *AMIA Annual Symposium Proceedings*, volume 2020, page 773. American Medical Informatics Association, 2020.

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

Namra Mahmood, Zahra Riaz, Arooj Sattar, and Mehwish Kiran. Hematological findings in covid-19 and their correlation with severity of disease. *Pakistan Journal of Medical Sciences*, 39(3):795, 2023.

Behrooz Mamandipoor, Wesley Yeung, Louis Agha-Mir-Salim, David J Stone, Venet Osmani, and Leo Anthony Celi. Prediction of blood lactate values in critically ill patients: a retrospective multi-center cohort study. *Journal of clinical monitoring and computing*, pages 1–11, 2022.

Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, and Xiangyang Ji. Supported trust region optimization for offline reinforcement learning. In *International Conference on Machine Learning*, pages 23829–23851. PMLR, 2023.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Normalizing flows for interventional density estimation. In *ICML*, 2023.

Sarah Müller, Alexander von Rohr, and Sebastian Trimpe. Local policy search with bayesian optimization. *Advances in Neural Information Processing Systems*, 34:20708–20720, 2021.

Mila Nambiar, Supriyo Ghosh, Priscilla Ong, Yu En Chan, Yong Mong Bee, and Pavitra Krishnaswamy. Deep offline reinforcement learning for real-world treatment optimization applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4673–4684, 2023.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Natalie M Pageler, Deborah Franzon, Christopher A Longhurst, Matthew Wood, Andrew Y Shin, Eloa S Adams, Eric Widen, and David N Cornfield. Embedding time-limited laboratory orders within computerized provider order entry reduces laboratory utilization. *Pediatric Critical Care Medicine*, 14(4):413–419, 2013.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

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

Noveen Sachdeva, Yi Su, and Thorsten Joachims. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 965–975, 2020.

Adam C Salisbury, Kimberly J Reid, Karen P Alexander, Frederick A Masoudi, Sue-Min Lai, Paul S Chan, Richard G Bach, Tracy Y Wang, John A Spertus, and Mikhail Kosiborod. Diagnostic blood loss from phlebotomy and hospital-acquired anemia during acute myocardial infarction. *Archives of internal medicine*, 171(18):1646–1653, 2011.

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Reliable off-policy learning for dosage combinations. *arXiv preprint arXiv:2305.19742*, 2023.

Mina S Sedrak, Mitesh S Patel, Justin B Ziemba, Dana Murray, Esther J Kim, C Jessica Dine, and Jennifer S Myers. Residents' self-report on why they order perceived unnecessary inpatient laboratory tests. *Journal of hospital medicine*, 11(12): 869–872, 2016.

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

Karina Spoyalo, Annie Lalande, Chantelle Rizan, Sophia Park, Janet Simons, Philip Dawe, Carl J Brown, Robert Lillywhite, and Andrea J MacNeill. Patient, hospital and environmental costs of unnecessary bloodwork: capturing the triple bottom line of inappropriate care in general surgery patients. *BMJ Open Quality*, 12(3):e002316, 2023.

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

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

Shengpu Tang, Maggie Makar, Michael Sjoding, Finale Doshi-Velez, and Jenna Wiens. Leveraging factored action spaces for efficient offline reinforcement learning in healthcare. *Advances in Neural Information Processing Systems*, 35:34272–34286, 2022.

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

Carl Van Walraven and C David Naylor. Do we know what inappropriate laboratory utilization is?: A systematic review of laboratory clinical audits. *Jama*, 280(6):550–558, 1998.

Arpana R Vidyarthi, Timothy Hamill, Adrienne L Green, Glenn Rosenbluth, and Robert B Baron. Changing resident test ordering behavior: a multilevel intervention to decrease laboratory utilization at an academic medical center. *American Journal of Medical Quality*, 30(1):81–87, 2015.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

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

Aaron Wilson, Alan Fern, and Prasad Tadepalli. Using trajectory data to improve bayesian optimization for reinforcement learning. *The Journal of Machine Learning Research*, 15(1):253–282, 2014.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Jialong Wu, Haixu Wu, Zihan Qiu, Jianmin Wang, and Mingsheng Long. Supported policy optimization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:31278–31291, 2022.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Haoran Xu, Louis Agha-Mir-Salim, Zachary O'Brien, Dora C Huang, Peiyao Li, Josep Gómez, Xiaoli Liu, Tongbo Liu, Wesley Yeung, Patrick Thoral, et al. Varying association of laboratory values with reference ranges and outcomes in critically ill patients: an analysis of data from five databases in four countries across asia, europe and north america. *BMJ Health & Care Informatics*, 28(1), 2021.

Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 4376–4396. PMLR, 2022.

Ming Zhi, Eric L Ding, Jesse Theisen-Toupal, Julia Whelan, and Ramy Arnaout. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. *PloS one*, 8(11):e78962, 2013.

Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pages 1719–1735. PMLR, 2021.

J. Zimmerman, M. Seneff, Xiaolu Sun, D. Wagner, and W. Knaus. Evaluating laboratory usage in the intensive care unit: patient and institutional characteristics that influence frequency of blood sampling., 1997.

Appendix 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 3.

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 3, 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 mortal-

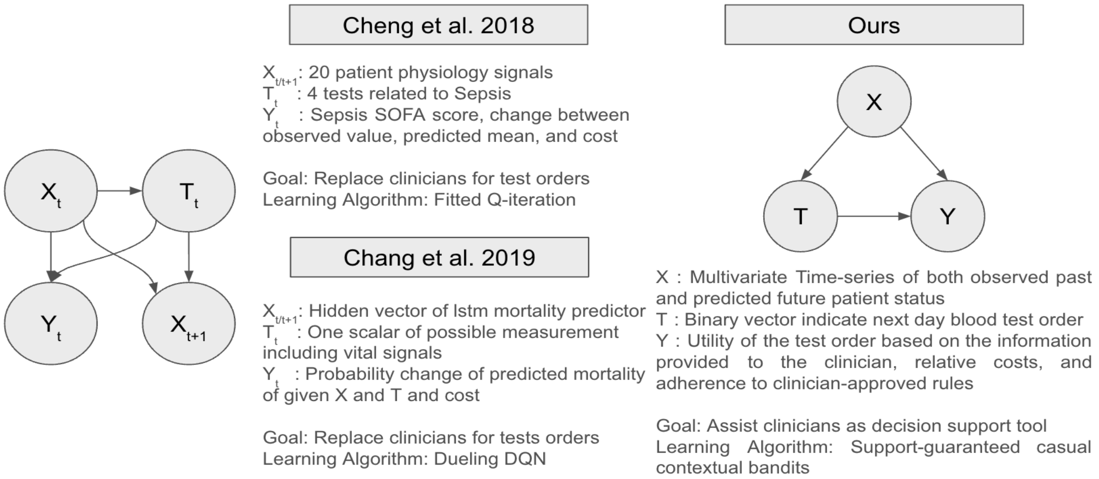
Cheng et al. 2018
X_{t/t+1}: 20 patient physiology signals
T_t : 4 tests related to Sepsis
Y_t : Sepsis SOFA score, change between observed value, predicted mean, and cost
Goal: Replace clinicians for test orders
Learning Algorithm: Fitted Q-iteration
Chang et al. 2019
X_{t/t+1}: Hidden vector of lstm mortality predictor
T_t : One scalar of possible measurement including vital signals
Y_t : Probability change of predicted mortality of given X and T and cost
Goal: Replace clinicians for tests orders
Learning Algorithm: Dueling DQN
X_t
T_t
Y_t
X_{t+1}
Ours
X
T
Y
X : Multivariate Time-series of both observed past and predicted future patient status
T : Binary vector indicate next day blood test order
Y : Utility of the test order based on the information provided to the clinician, relative costs, and adherence to clinician-approved rules
Goal: Assist clinicians as decision support tool
Learning Algorithm: Support-guaranteed casual contextual bandits

Figure 3: Graphical representation of two prior works (Cheng et al., 2018; Chang et al., 2019) and our proposed methods.

ity 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%.

A.1. Discussion of Ignorability Assumption

To uphold our ignorability assumption, we consider our covariate set, which includes treatments (for example, drugs and procedures), vital signals, and test results, to be complete. In other words, any drugs given in the past but not captured in these covariates are assumed to have no effect on lab test orders. If our covariates were to miss essential information for determining lab test orders, the ignorability assumption would no longer hold because unmeasured confounders would be present.

When might this assumption fail? In cases where a drug's effect is both intense at a single time point and expected to extend over multiple future steps, the bandit framework may not fully capture its impact. This naturally leads us to consider strategies for relaxing this assumption.

One approach is to explicitly relax the assumption by incorporating the dynamic effects of patient covariates—including drug prescriptions—on lab test ordering, effectively framing the problem as an offline reinforcement learning task within a Markov Decision Process. While this extension could broaden EXOSITO's application to various diseases, we believe the bandit model is well-suited for the ICU context, where patient biomarkers are monitored and treatments are frequently adjusted. Alternatively, a simpler solution is to adjust the covariates so that the extended effects of drugs are represented over time while still using the bandit framework, thereby ensuring that long-lasting drug influences are considered when determining lab test orders.

Appendix 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 and 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 preci-

Table 3: 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

Table 4: Test set performance for trained time-series forecasting models. We found that the PatchTST model obtained the lowest forecasting error and formed the backbone for our forecasting module

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**

sion. 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 4.

B.1. Evaluating Patient Status Forecasting Model

As mentioned in the discussion session, the performance of the time-series forecasting model can impact the subsequently learned policy. To better clarify this point, we conducted a new evaluation of our time-series predictor on the test set, assessing performance (MSE) by calculating the difference at each time point (each hour of the predicted future 24 hours) and for each feature. Our analysis verified that across all time points, the MSE consistently hovers around 0.03, which aligns with our reported results.

To gain further insights, we evaluated the MSE for each feature across all test set samples and grouped features into three categories based on their average MSE:

- Low MSE Group

  Features (Counts): CreatinineKinase (6977), MRI (931), PronePosition (638), Vasopressors (86261), Antipsychotics (1924), TPN (1283), UltraSound (1555), EnteralNutrition (1844), CTScan (2998), Antiarrhythmics (2705), PReplacement (2199), Paralysis (2591), Antiepileptics (4210), ICPMonitor (21547), Troponin (5656), SAS (40291), Calcium (53351), Lactate (32035), MgReplacement (10866), Sodium (58406), Magnesium (57475)

  Average MSE: 0.0059

- Medium MSE Group

  Features (Counts): TidalVolume (79486), Hemoglobin (54710), PaCO2 (51837), Bicarbonate (57018), Xray (12783), ALP (14574), Transfusions (10977), FiO2 (115116), ph (52246), ICDSC (32044), WBC (52127), Bilirubin (14774), Potassium (72510), AirwayPressure (80863), INR (36491), ALT (14720), Anticoagulants (18484), Phosphate (53024), Diuretics (16520), PPI (18118), PEEP (86456)

  Average MSE: 0.0195

- High MSE Group

  Features (Counts): Platelets (52386), CaReplacement (13279), UrineOutput (478588), Antihypertensives (38794), MinuteVentilation (81693), BloodUreaNitrogen (59222), Dialysis (30275), Temperature (213724), PaO2 (44297), Creatinine (58271), GCS (227697), KReplacement (31494), Analgesia (70894), Antibiotics (70132), Sedation (83964), DiastolicBloodPressure (881326), MeanBloodPressure (881395), HeartRate (917866), SystolicBloodPressure (882387), RespiratoryRate (909106), Ventilation (88295)

  Average MSE: 0.0801

We observed that vital signals (e.g., heart rate), which require less manual effort to collect, are more frequently represented in the data, whereas treatments and lab values are more sparsely recorded, reflecting the real-time nature of patient monitoring. Despite these challenges, we believe incorporating the forecasting model into our approach enhances the ability of the learned policy to make clinically relevant lab order recommendations based on both present and (predicted) future for patient status.

Appendix 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 who are senior attending physicians in leading hospitals together with support form peer-reviewed literature (Kumwilaisak et al., 2008; Cismondi et al., 2013; Vidyarthi et al., 2015; Bindraban et al., 2018), 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 4, 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.

Appendix D. Design of ΔX

During consultations with practitioners, clinicians noted that most patients tend to have consistent lab results, which are often classified as 'abnormal.' However, our clinical collaborators emphasized that it is more informative to observe extreme or mean changes in lab results. The motivation for designing ΔX to measure the range or mean changes is that clinicians are more likely to order a test if the test result is expected to change significantly (either improving or deteriorating).

As introduced in Section 4.3, our framework assumes that lab test orders are more informative to clinicians when there are larger changes in lab values—whether these are extreme changes or significant shifts in mean values. ΔX is specifically designed to encourage the policy to prioritize tests that are likely to provide such informative insights.

The design choice of ΔX is supported both by our conversations with clinical collaborators and prior literature.

Çakırca et al. (2023) and Mahmood et al. (2023) both provide a comparative analysis of blood test abnormalities in ICU vs. non-ICU COVID-19 patients. The results show that WBC are remarkably higher in the ICU patients than non-ICU patients while albumin levels are remarkably lower for ICU patients compared to non-ICU patients. The authors concluded that ICU patients' conditions are more severe than the general patient population outside the ICU.

In Xu et al. (2021), the authors specifically compared lab results of creatinine, hemoglobin, lactate, sodium and showed that ICU patients normally have a much higher variance in these test results and used reference ranges are larger than non-ICU patients. This study suggests that standard reference ranges have limited relevance for ICU patients and highlights the need for context-specific ranges to enhance clinical interpretation.

The ICU-Labome study Alkozai et al. (2018) evaluated 35 routine laboratory measurements in over 49,000 ICU patients. The research found that many laboratory values were outside standard reference intervals. Notably, 14 out of 35 measurements had median values outside standard reference intervals, underscoring the necessity for ICU-specific reference ranges.

Authors in Mamandipoor et al. (2022) build machine learning models that predict lactate values for ICU patients and show that elevations in serum lactate levels are strong predictors of mortality. Thus, anticipating these changes allows clinicians to intensify care proactively. In addition, ordering routine tests for ICU patients is common medical practice. However, studies also suggest that continuing to order certain lab tests when their values remain stable may not be beneficial and can even be harmful (Kleinpell et al., 2018; Eaton et al., 2017; Levi et al., 2019).

Kleinpell et al. (2018) recommend against routine daily laboratory tests for clinically stable ICU patients, as such practices often do not contribute to patient care and can lead to unnecessary interventions. Excessive blood draws for stable ICU patients can lead to hospital-acquired anemia, increased healthcare costs, and unnecessary downstream testing and procedures (Eaton et al., 2017). Implementing strategies to reduce unnecessary laboratory testing in the ICU has been shown to decrease the volume of blood collected per patient-day without negatively affecting patient outcomes (Levi et al., 2019).

The references provided support the understanding that ICU patients generally exhibit more consistently abnormal laboratory values compared to patients in other wards. Building on this, we argue that instead of concentrating on lab values that are stable yet abnormal, it would be more clinically valuable to focus on identifying and prioritizing tests for lab values that demonstrate significant changes over time. These dynamic changes are more likely to provide actionable insights and support timely clinical interventions.

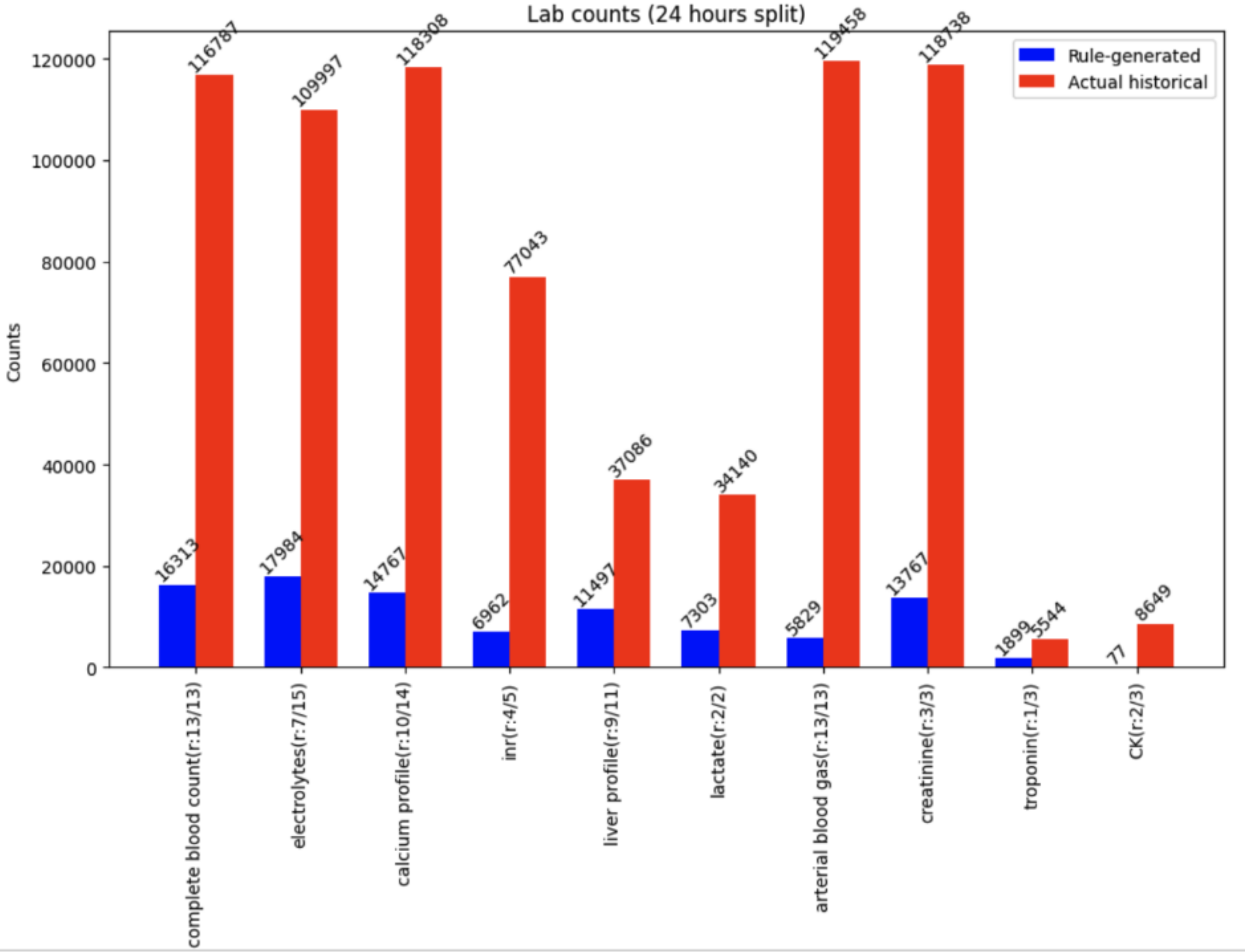

Figure 4: Bar plot that shows the distribution of guideline-generated tests and observed tests

Appendix E. 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 and Mohamed (2015); Tabak and 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.

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 and 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 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

Appendix F. Policy Learning Algorithm and Hyperparameters for Policy Training

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:

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

where $\Pi^{\text{r}} = \left\{ \pi \in \Pi \mid \hat{f}\left(\pi(x), x\right) > \bar{\varepsilon} \right\},\ \forall x \in \mathcal{X}\}$ 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 utility 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.(6) 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. (7), 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.

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. (7), 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. Our method is trained with NVIDIA A6000 GPUs, one single A6000 GPU would be able to complete the training.

For evaluation, we slightly modified our lab test order utility function $g(t, x)$ to suit as a metric. During testing, we set $\beta_1 = \beta_2 = 1$ for outcome calculation in Eq. (5). Additionally, we adapted the smooth term $\mathcal{L}_b$ from Eq. (3) to a discrete form: $\mathcal{L}_b^{test}(t, x) = L_{low} + L_{up}$, where $L_{low} = \sum_{j=1}^{K} \mathbb{1}(t_j < 0.5, t_j^{low} = 1)$ and $L_{up} = \sum_{j=1}^{K} \mathbb{1}(t_j > 0.5, t_j^{up} = 0)$. 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.

Appendix G. Additional Experiments on MIMIC and HIRID Dataset

We conducted further experiments on the MIMIC dataset, adjusting for the minimum required lab tests ordered in each data point. These experiments consistently showed that our method, employing GPS-guided policy learning, outperforms those based on physician decisions and conventional RL approaches, with mortality as the reward metric.

In addition to our work on the MIMIC dataset, we aimed to validate the applicability of our method on a broader scale. We analyzed a recently released ICU dataset from patients in a Swiss hospital, known as the High Time Resolution ICU dataset (HiRID) Hyland et al. (2020). This dataset, initially utilized to predict circulatory failure, has mostly been explored through its inferred version in previous studies. However, our detailed examination of HiRID's raw data revealed that it serves as an apt irregular time-series dataset for our objectives, similar in structure to MIMIC. We present an overview of this newly processed HiRID dataset alongside the results of applying our method, mirroring the experimental approach taken with the MIMIC dataset.

G.1. Various Patient Time-series Covariates Settings

Our method begins by constructing a time-series forecasting model, ϕ, that predicts the future ICU stay status of a patient based on observed data. Among the 71 features in our covariates X, 21 are related to blood test values. To validate the trained model, our main experiment ensured that for each 24-hour period, at least 5 tests were ordered, meaning X_{prev}, covering 48 hours, should include results from at least 10 tests, and X_{post}, spanning 24 hours, from at least 5 tests. This requirement explains why the results in Table 5 show an average of 6 to 8 tests ordered for the next day.

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

Policy	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
Random$_{0.5}$	0.23 ±0.01	0.51±0.01	4.3±0.01	3.2±0.01	1.1±0.01
Random$_{0.75}$	0.34 ±0.01	0.75±0.01	3.64±0.01	1.6±0.01	2.04±0.01
LowerBound	0.37	0.62	0	0	0
UpperBound	0.44	0.82	0	0	0
Physician	0.41	0.67	1.24	1.24	0
Ours(w/o GPS)	0.44 ±0.01	0.8 ±0.003	1.06±0.001	0.34	0.72
Ours(w GPS)	0.42 ±0.005	0.66±0.005	1.16±0.001	0.67	0.49

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

Policy	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	info gain↑
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±0.01	0.8 ±0.003	1.06±0.001	(1.2)
Ours(w GPS)	0.42±0.005	0.66±0.005	1.16 ±0.001	(0.98)

To prepare for eventual deployment in ICU wards, we tested various settings with even fewer minimum required data points in X, aiming to more closely mirror real-world situations where some patients may not undergo any tests within a 24-hour period.

With this consideration, we established two additional settings for the minimum number of required tests: the first requires X_{prev} to include results from at least 5 lab tests and X_{post} from at least 3 tests; the second mandates X_{prev} to contain results from at least 1 lab test, with no requirements for X_{post}. Despite these adjustments, we ensured that each data point included at least 15 non-missing entries collected over the patient's 48-hour ICU stay of interest.

As we reduced the minimum number of required tests, we observed an increase in the number of zeros or missing values in the data points. For our PatchTST forecasting model, the loss decreased as the minimum number of required tests was lowered. Given that PatchTST outperformed other models in our forecasting experiments, we evaluated its performance under these three settings of minimum test requirements. The findings are detailed in Table 7.

Subsequently, we employed these two additional versions of ϕ to explore the learning of our lab test ordering policy, guided by our proposed outcome function $g(t, x)$.

G.2. Consistent MIMIC Results for Different Covariates Settings

Incorporating two additional settings for the minimum required number of tests, we extended our experiments as detailed in the results section of the paper. Table 8 displays the test set performance for both baseline models and our learned policies with a minimum of 5 tests required every 24 hours. Table 9 presents the performance for baseline models and our learned policies with only 1 minimum test required every 48 hours.

Firstly, across these settings, we observed a consistent trend: our learned policy, supported by the propensity score function, consistently orders tests that not only provide clinicians with more relevant information but also generate lower costs and adhere more closely to clinical rules, thereby missing fewer necessary tests compared to the physician's policy.

A notable decrease in costs across methods, except for the random policy, was observed. This reduction is attributed to the less frequent testing of patient ICU stays due to the altered minimum test requirements. As evidenced in Table 9, an average of 2 to 3 tests are ordered every 24 hours. Reducing the number of minimally required tests led to an increase in L_{up}, indicating a decrease in the maximum number of tests that should be ordered, and a tendency for learned policies to order more tests than necessary. Nonetheless, guided by our outcome function

Table 7: The test set performance for trained PatchTST time-series forecasting model with different minimum number of ordered tests

Number of minimum test required	MSE	MAE
10, 5 (Main result)	0.027± 0.001	0.059± 0.002
5, 3	0.024± 0.002	0.054± 0.003
1, 0	0.018± 0.002	0.038± 0.001

Table 8: Testset performance for baseline and our learned policies (with std). Minimum required tests is 5 per 48 hours.

Policy	$\Delta X \uparrow$	Cost↓	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
Random$_{0.5}$	0.21 ±0.01	0.51±0.01	4.3±0.01	3.2±0.01	2.1±0.01
LowerBound	0.13	0.24	0	0	0
UpperBound	0.25	0.47	0	0	0
Physician	0.17	0.36	1.18	1.18	0
Ours(w/o GPS)	0.24 ±0.007	0.46 ±0.003	0.94 ±0.003	0.13	0.81
Ours(w GPS)	0.2 ±0.002	0.32±0.01	1.09±0.004	0.4	0.69

$g(t, x)$, our method consistently achieved the best $\mathcal{L}_b$ across all approaches. Moreover, the similarities between the numbers in Table 8 and Table 9 suggest that most ICU stays in the MIMIC dataset typically include around 5 test data points every 48 hours.

Furthermore, Table 9 highlights our method's superiority in conditions more closely mirroring real-world scenarios compared to the traditional physician policy (logging policy), underscoring our method's effectiveness.

Additionally, we compared our approach to an RL method that uses mortality as the reward. Under more realistic settings, the RL method showed diminished performance in terms of cost, clinical relevancy, and utility to clinicians. Moreover, our approach, which does not rely on mortality as a learning signal, achieved comparable or even superior information gain relative to the RL method focused on patient mortality. This comparison further underscores the importance of designing ICU lab test ordering systems that prioritize clinician needs over patient-facing metrics.

G.3. HIRID Results

The High Time Resolution ICU Dataset (HiRID) Hyland et al. (2020) is a publicly available critical care dataset that originates from a collaboration between the Swiss Federal Institute of Technology (ETH Zurich) and the University Hospital of Bern. HiRID contains a rich collection of high-resolution data from patients admitted to the intensive care unit (ICU), designed to support a wide range of research initiatives in critical care medicine and machine learning.

Spanning over several years, HiRID includes data from thousands of ICU stays, offering detailed information on physiological parameters, laboratory test results, treatment interventions, and more. One of the dataset's distinguishing features is its high temporal resolution, providing minute-by-minute measurements for a subset of variables, which enables the development and validation of predictive models that require fine-grained temporal data.

Originally developed to facilitate the prediction of circulatory failure and other critical events in the ICU, HiRID's comprehensive and detailed nature makes it suitable for a broad array of research questions. This includes studies on disease progression, treatment effect analysis, and the development of decision support tools for clinicians. The dataset's structure allows for the exploration of irregular time-series data in medical contexts, making it an invaluable resource for advancing patient care through machine learning and data-driven approaches.

While the MIMIC dataset provides a substantial repository with approximately 57,000 ICU stays, the

Table 9: Testset performance for baseline and our learned policies (with std). Minimum required tests is 1 per 48 hours.

Policy	$\Delta X \uparrow$	Cost$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
Random$_{0.5}$	0.21 ±0.01	0.51±0.01	4.2±0.01	3.1±0.01	2.1±0.01
LowerBound	0.09	0.18	0	0	0
UpperBound	0.23	0.41	0	0	0
Physician	0.15	0.33	1.12	1.12	0
Ours(w/o GPS)	0.21 ±0.001	0.38 ±0.002	0.91 ±0.003	0.07	0.84
Ours(w GPS)	0.18 ±0.001	0.28±0.01	1.03±0.004	0.22	0.79

Table 10: Testset performance of prior work and ours policies (with std) with 1 minimum required tests per 48 hours.

Policy	$\Delta X \uparrow$	Cost$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	info gain$\uparrow$
Physician	0.15	0.33	1.12	1.07
RL (low cost)	-	0.32	2.1	1
RL (high cost)	-	0.78	5.3	1.24
Ours(w/o GPS)	0.21 ±0.001	0.38 ±0.002	0.91 ±0.003	(1.22)
Ours(w GPS)	0.18 ±0.001	0.28±0.01	1.03±0.004	(1.09)

High Time Resolution ICU Dataset (HiRID) encompasses around 32,000 stays. Despite this smaller number, HiRID compensates with its detailed and frequent collection of patient data, especially concerning vital signs and treatment interventions. This high granularity makes HiRID an exemplary dataset for irregular time-series analysis, offering a dense array of data points and significantly less missingness, which is instrumental in the development of advanced patient forecasting models.

Upon processing the HiRID dataset, we identified a total of 79 features, including a more diverse assortment of Vasopressors compared to those found within the MIMIC dataset. This diversification in data points allows for a more nuanced analysis of patient responses to various treatments. Notably, the mean squared error for HiRID stands at 0.035 ± 0.003, a figure that, while higher than that observed in MIMIC, reflects the reduced incidence of missing data within the dataset. Despite this increased error margin, the loss remains commendably low given the sparsity of the data, suggesting that forecasting models ϕ developed using HiRID data possess a higher degree of reliability and applicability than those trained exclusively on MIMIC data.

In applying our machine learning methodology to the HiRID dataset, we followed the same procedural rigor as with our MIMIC dataset application. The results from this endeavor align closely with our baseline comparisons, underscoring the robustness and efficacy of our approach as detailed in Tables 11 and 12. This consistency across datasets reinforces the validity of our method and its potential for real-world clinical application.

Notwithstanding the successful application and experimentation with HiRID—a dataset derived from a separate patient cohort—our investigation into the robustness of our methodology in the face of data distribution shifts remains in its infancy. The primary challenge lies in reconciling the differing nomenclatures and feature combinations present in each dataset, such as the variations in Antibiotics listed across MIMIC and HiRID. Nevertheless, the ability to corroborate our findings with an additional, independently collected dataset bolsters confidence in the generalizability and practical applicability of our method.

Looking ahead, these preliminary findings lay the groundwork for more exhaustive future investigations into the interoperability of models trained on one dataset and tested on another, particularly between MIMIC and HiRID. Such research will be crucial in assessing the adaptability and versatility of our machine learning strategies across varied clinical datasets, marking a promising direction for subsequent work.

Appendix H. 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.

H.1. 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 utility function $g(t, x)$ and in the bound loss, particularly when the forecasted future is excluded from policy training.

H.2. 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. These numbers are suggested by clinicians with prior studies Kandalam et al. (2020); Spoyalo et al. (2023). 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.

H.3. 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.

Table 11: Testset performance for baseline and our learned policies (with std), on HiRID dataset.

Policy	$\Delta X \uparrow$	Cost$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	$L_{low} \downarrow$	$L_{up} \downarrow$
Random$_{0.5}$	0.47 ±0.01	0.49±0.1	4.4±0.01	2.9±0.01	1.5±0.03
Random$_{0.75}$	0.58 ±0.01	0.74±0.1	3.77±0.01	1.66±0.01	2.11±0.01
LowerBound	0.62	0.35	0	0	0
UpperBound	1.13	0.59	0	0	0
Physician	0.96	0.55	0.98	0.98	0
Ours(w/o GPS)	1.08 ±0.01	0.57 ±0.02	0.62±0.001	0.3	0.32
Ours(w GPS)	1.01 ±0.01	0.52±0.01	0.89±0.001	0.5	0.39

Table 12: Testset performance of prior work and ours policies (with std), on HiRID datset.

Policy	$\Delta X \uparrow$	Cost$\downarrow$	$\mathcal{L}_b^{test} \downarrow$	info gain$\uparrow$
Physician	0.96	0.55	0.98	1.03
RL (low cost)	-	0.51	1.4	1
RL (high cost)	-	0.74	3.6	1.19
Ours(w/o GPS)	1.08 ±0.01	0.57 ±0.02	0.62±0.001	(1.17)
Ours(w GPS)	1.01 ±0.01	0.52±0.01	0.89±0.001	(1.05)

H.4. Ablation Study on Time-series Model with Different Level of Errors

We want to investigate how prediction errors in the time-series model influence the learned policy. To answer this, we ran two ablation studies to test how the time-series model error would affect the evaluation results (reward function g):

- For each test sample, we choose to generate the predicted next 24 hours either by random OR using our learned time-series model. Then the policy function is provided input with either [48-hour observation, random next 24-hour] or [48-hour observation, learned model predicted next 24-hour]. Using random predictions resulted in the cost metric for the next 24 hours going up 50%, L_b went up 43% with delta X went up only 3%. The big increase of cost and L_b means that given the random prediction as input, our learned policy is ordering more lab tests that deviate from rule-generated tests or the physician (logging) policy shown in the dataset. However, even though more tests are being ordered the information provided to the clinical team was not significantly increased (as indicated by the small increase in ΔX).
- We also compared the actions generated with input of the true next 24-hour and the orders generated using the learned time-series model predicted 24-hour. When using the true next 24-hour as input, our policy generates actions that have a 3% decrease in cost, 6% decrease in L_b and 5% increase in ΔX. The small increase in cost and L_b shows the perfect prediction input allows our learned policy to order a bit more tests to provide a little more information (increase in ΔX) to the clinicians.

This shows that our learned time-series model is doing a solid but not perfect job on predicting the patient status since using evaluating our learned policy with perfect prediction is not significantly different from evaluating our learned policy with our learned times-series model prediction.

In conclusion, if the prediction error of the time-series model is high, then the outcome/reward function will show a worse reward than the time-series model with less prediction error.

In part, our utility of this framework does derive from the quality of the time-series prediction and beyond simplistic settings, it is difficult to perfectly characterize the relationship between the two.

H.5. Ablation Study on Simulating 'Human Error' in the Privileged Information

We investigate how errors in rule specification affect the policy learning. In our framework, the rules serve as indicators or bounds that define the minimal set of lab tests to be ordered based on the patient's status. These rules are not meant to be exhaustive or flawless but provide a conservative baseline to ensure safe practice.

Consider two extremes: if we had perfect rules that match the Bayes optimal predictor for the policy function, we could directly solve the problem using those rules. In contrast, if no rules are applied, the policy's search space ranges from ordering no tests at all to ordering all tests observed under the logging policy; in Algorithm 1, this scenario corresponds to having all t^{lower} vectors set to zero. Any inclusion of rules results in non-zero t^{lower}, which restricts the search space and guarantees that certain tests are ordered to mitigate the risk of missing necessary tests.

The above argument suggests that there is a statistical benefit to having rules. So what happens if the rules contain human errors.

We first define that we say a rule is more conservative if the condition of the rule is more extreme (less likely to trigger the rule and order corresponding tests). In other words, a more conservative rule will yield more 0's in t^{lower} than a less conservative rule.

The human error in the rules therefore yields more conservative rules or less conservative rules.

If human error leads to more conservative rules (i.e. all zeros in t^{lower}), this is equivalent to not including any rules in our framework, which means our policy is solving causal bandits without side information.If human error leads to less conservative rules (i.e. more 1's in t^{lower}), it would result in over-ordering as in the logging policy.

On the flipside, there are two possibilities for less conservative rules. a] Either a less conservative rule would lead to t^{lower} becoming a vector of 1's, meaning each patient should get every test order after each 48 hours. But this implies that the physicians in the logging policy (current medical practice) are under-ordering lab tests for patients, which contradicts our assumption that logging policy has over-ordering issues and is suboptimal.

b] Or that less conservative rules would lead to t^{lower} having similar 1's as t^*, means that the clinician defined rules are ordering similar amounts of lab tests as the logging policy. This also implies that the rules recover a similar ordering policy as the physician policy.

To further support the above argument, we perform another ablation study on the rules. We randomly choose three lab tests from our study and remove all the rules related to these three lab tests, meaning t^{lower} will always be 0 for these three lab tests. We trained the policy with our framework and during evaluation, we found that the total number of these three tests ordered is roughly 9% less than training with their corresponding rules due to the goal of minimizing cost in the reward. This result shows that if we increase the search space for lab orders, with the guide of our reward function, lab tests that are less triggered by each of these rules would be ordered less. This leads to a policy that takes more risks to potentially miss necessary tests.

With the same three lab tests, instead of removing all corresponding rules, we change the rules to 'suggest to order if the dataset is ordering', meaning that t^{lower} is always equal to t^{upper} for these three lab tests. During evaluation, we found that the total number of these three test orders increased 37% compared to training with our existing rules and the total number of lab orders for other tests decreased 23%. This result indicates that if we decrease the search space for lab orders, with the guide of our reward function, lab tests that are triggered by rules similar to the logging policy would be ordered more and cause over-ordering to happen.

Thus, having rules help the policy generate minimal lab orders and having less conservative rules would revert the policy back to the logging policy which errs on over-ordering.

H.6. Ablation Study on the Use of Privileged Information (Clinical Rules)

We assessed the impact of incorporating privileged information—specifically, clinical rules—into our policy learning framework. These rules serve as guidance to ensure safety and informativeness of the lab test orders selected by the learned policy. We performed an ablation study by comparing our original approach (using Algorithm 1 to compute minimal necessary tests, t_{lower}) against two altered conditions:

1. No Rules: Instead of using clinical rules to determine t_lower, we set t_lower to zero vectors, effectively removing any privileged guidance.

2. Rules as the Logging (Physician) Policy: Instead of using clinical rules, we force t_lower to be the same as the logging policy. This simulates having rules identical to the logging policy, which may over-order tests.

Setting 1 (No Rules):

Behavior of the Learned Policy: Without any privileged information, the policy converged to ordering only low-cost tests (e.g., CK, Creatinine) and avoided relatively expensive tests (e.g., ABG, INR, Troponin). This approach effectively minimizes the cost penalty but disregards safety and informativeness.

Performance Metrics:

- L_{upper} increased by about 11%: The policy's test orders deviate significantly from the physician's (logging) policy and the clinical rules.
- L_{lower} increased by about 56%: The learned policy's behavior diverges greatly from the clinical standard, indicating poor adherence to necessary tests.
- Cost metric unchanged: The number of tests ordered does not decrease, but the composition shifts toward cheaper tests.
- ΔX decreased by about 7%: The selected tests provide less informative clinical insight.

These results suggest that without privileged rules, the policy exploits the cost structure and may adopt a clinically suboptimal strategy. Further adjusting the cost weight could lead to trivial (zero-test) solutions, which are even less clinically informative.

Setting 2 (Rules = Logging Policy):

Behavior of the Learned Policy: By setting the rules to match the logging physician policy, the learned policy closely mimics the logging policy's test selection, leading to negligible differences in L_{upper}.

Performance Metrics:

- L_{lower} increased by about 15% compared to the logging policy and by about 34% compared to using clinical rules. Although the learned policy adheres to the logging pattern, it misses critical tests that clinical rules would have suggested.
- Cost decreased by about 6% compared to the policy trained with clinical rules. This indicates fewer tests being ordered overall.
- ΔX decreased by about 17%, showing that the selected tests are less informative to the clinical team.

These findings show that using the logging policy as "rules" constrains the policy to the baseline ordering pattern, offering limited incentive to discover more appropriate, informative, and cost-effective test combinations.

The ablation study demonstrates that having privileged clinical rules is critical. Without these rules (Setting 1), the policy finds cost-minimizing but clinically uninformative solutions. Using the logging policy as the rule-set (Setting 2) leads to overfitting on the baseline pattern and fails to discover safer and more informative sets of tests. In contrast, integrating clinical rules (privileged information) ensures that the learned policy can explore a broader range of test combinations while maintaining a clinically safe and informative standard.

Appendix I. Method Explanation for Non-ML Audience

In our work, we've developed a machine learning method aimed at optimizing blood test ordering for ICU patients, making this process more efficient and informed by data. Our goal is to explain this approach in straightforward terms, particularly for clinicians who could integrate this tool into their daily routines, enhancing patient care without getting bogged down by complex mathematical formulas.

Our method uses existing patient data from Electronic Health Records (EHR) to create two key tools: a forecasting model (represented by ϕ) and a decision support policy (π). The forecasting model analyzes the past 48 hours of a patient's data to predict their health status over the next 24 hours, including lab results, treatment responses, and vital signs. The decision support policy then uses this prediction to recommend which blood tests should be ordered for the next day for each patient.

Imagine integrating this tool into an ICU setting. The decision support tool, combined with the forecasting model (ϕ), guides clinicians on which tests to order next. For example, if the policy (π) suggests ordering a Complete Blood Count (CBC), a clinician might wonder why. By consulting the model's predictions or the patient's recent data, the clinician can see that the patient recently had a blood transfusion, justifying the need for a CBC test. This approach, recommended by our clinical partners, moves away from abstract data representations and instead bases decisions on actual patient data, making the reasoning behind each test clear.

The significant advantage of deploying our method is its ability to recommend a focused set of blood tests that are both cost-effective and clinically relevant, avoiding unnecessary tests while adhering to clinical guidelines. This not only has the potential to lighten the workload for hospital staff but also ensures patient safety is not compromised.

By directly leveraging real patient data for both present and future predictions, our system offers a clear, rational basis for each recommended test, aligning closely with clinical needs and practices. This could revolutionize how clinicians make decisions about patient care, making the process more efficient and grounded in data-driven insight