An Optimal Policy for Patient Laboratory Tests in Intensive Care Units

Li-Fang Cheng*[1], Niranjani Prasad*[2] and Barbara E. Engelhardt[2,3]

[1] *Department of Electrical Engineering, Princeton University*
[2] *Department of Computer Science, Princeton University*
[3] *Center for Statistics and Machine Learning, Princeton University*

Laboratory testing is an integral tool in the management of patient care in hospitals, particularly in intensive care units (ICUs). There exists an inherent trade-off in the selection and timing of lab tests between considerations of the expected utility in clinical decision-making of a given test at a specific time, and the associated cost or risk it poses to the patient. In this work, we introduce a framework that learns policies for ordering lab tests which optimizes for this trade-off. Our approach uses batch off-policy reinforcement learning with a composite reward function based on clinical imperatives, applied to data that include examples of clinicians ordering labs for patients. To this end, we develop and extend principles of Pareto optimality to improve the selection of actions based on multiple reward function components while respecting typical procedural considerations and prioritization of clinical goals in the ICU. Our experiments show that we can estimate a policy that reduces the frequency of lab tests and optimizes timing to minimize information redundancy. We also find that the estimated policies typically suggest ordering lab tests well ahead of critical onsets—such as mechanical ventilation or dialysis—that depend on the lab results. We evaluate our approach by quantifying how these policies may initiate earlier onset of treatment.

Keywords: Reinforcement Learning, Dynamic Treatment Regimes, Pareto Learning

1. Introduction

Precise, targeted patient monitoring is central to improving treatment in an ICU, allowing clinicians to detect changes in patient state and to intervene promptly and only when necessary. While basic physiological parameters that can be monitored bedside (e.g., heart rate) are recorded continually, those that require invasive or expensive laboratory tests (e.g., white blood cell counts) are more intermittently sampled. These lab tests are estimated to influence up to 70% percent of diagnoses or treatment decisions, and are often cited as the motivation for more costly downstream care [1, 2]. Recent medical reviews raise several concerns about the over-ordering of lab tests in the ICU [3]. Redundant testing can occur when labs are ordered by multiple clinicians treating the same patient or when recurring orders are placed without reassessment of clinical necessity. Many of these orders occur at time intervals that are unlikely to include a clinically relevant change or when large panel testing is repeated to detect a change in a small subset of analyses [4]. This leads to inflation in costs of care and in the likelihood of false positives in diagnostics, and also causes unnecessary discomfort to the patient.

*These authors contributed equally to this work.

Moreover, excessive phlebotomies (blood tests) can contribute to risk of hospital-acquired anaemia; around 95% of patients in the ICU have below normal haemoglobin levels by day 3 of admission and are in need of blood transfusions. It has been shown that phlebotomy accounts for almost half the variation in the amount of blood transfused [5].

With the disproportionate rise in lab costs relative to medical activity in recent years, there is a pressing need for a sustainable approach to test ordering. A variety of approaches have been considered to this end, including restrictions on the minimum time interval between tests or the total number of tests ordered per week. More data-driven approaches include an information theoretic framework to analyze the amount of novel information in each ICU lab test by computing conditional entropy and quantifying the decrease in novel information of a test over the first three days of an admission [6].

In a similar vein, a binary classifier was trained using fuzzy modeling to determine whether or not a given lab test contributes to information gain in the clinical management of patients with gastrointestinal bleeding [7]. An "informative" lab test is one in which there is significant change in the value of the tested parameter, or where values were beyond certain clinically defined thresholds; the results suggest a 50% reduction in lab tests compared with observed behaviour. More recent work looked at predicting the results of ferratin testing for iron deficiency from information in other labs performed concurrently [8]. The predictability of the measurement is inversely proportional to the novel information in the test. These past approaches underscore the high levels of redundancy that arise from current practice. However, there are many key clinical factors that have not been previously accounted for, such as the low-cost predictive information available from vital signs, causal connection of clinical interventions with test results, and the relative costs associated with ordering tests.

In this work, we introduce a reinforcement learning (RL) based method to tackle the problem of developing a policy to perform actionable lab testing in ICU patients. Our approach is two-fold: first, we build an interpretable model to forecast future patient states based on past observations, including uncertainty quantification. We adapt multi-output Gaussian processes (MOGPs; [9, 10]) to learn the patient state transition dynamics from a patient cohort including sparse and irregularly sampled medical time series data, and to predict future states of a given patient trajectory. Second, we model patient trajectories as a Markov decision process (MDP). This framework has been applied to the recommendation of treatment strategies for critical care patients in a variety of different settings, from recommending drug dosages to efficiently weaning patients from mechanical ventilation [11–13]. We design the state and reward functions of the MDP to incorporate relevant clinical information, such as the expected information gain, administered interventions, and costs of actions (here, ordering a lab test). A major challenge is designing a reward function that can trade off multiple, often opposing, objectives. There has been initial work on extending the MDP framework to composite reward functions [14]. Specifically, fitted Q-iteration (FQI) has been used to learn policies for multi-objective MDPs with vector-valued rewards, for the sequence of interventions in two-stage clinical antipsychotic trials [15]. A variation of Pareto domination was then used to generate a partial ordering of policies and extract all policies that are optimal for some scalarization function, leaving the choice of parameters of the scalarization function to decision makers.

Here, we look to translate these principles to the problem of lab test ordering. Specifically, we focus on blood tests relevant in the diagnosis of sepsis or acute renal failure, two common conditions associated with high mortality risk in the ICU: white blood cell count (WBC), blood lactate level, serum creatinine, and blood urea nitrogen (BUN). We present our methods within a flexible framework that can in principle be adapted to a patient cohort with different diagnoses or treatment objectives, influenced by a distinct set of lab results. Our proposed framework integrates prior work on off-policy RL and Pareto learning with practical clinical constraints to yield policies that are close to intuition demonstrated in historical data. We apply our framework to a publicly available database of ICU admissions, evaluating the estimated policy against the policy followed by clinicians using both importance sampling based estimators for off-policy policy evaluation and by comparing against multiple clinically inspired objectives, including onset of clinical treatment that was motivated by the lab results.

2. Methods

2.1. *Cohort selection and preprocessing*

We extract our cohort of interest from the MIMIC III database [16], which includes de-identified critical care data from over 58,000 hospital admissions. From this database, we first select adult patients with at least one recorded measure for each of 20 vital signs and lab tests commonly ordered and reviewed by clinicians (for instance, results reported in a complete blood count or basic metabolic panel). We further filter patients by their length-of-stay, keeping only those in the ICU for between one and twenty days, to obtain a final set of 6,060 patients (Table 1).

Table 1. **Total recordings, mean & standard deviation (SD)** for covariates in selected cohort.

Covariate	Count	Mean	SD
Respiratory Rate (RR)	1,046,364	20.1	5.7
Heart Rate (HR)	964,804	87.5	18.2
Mean Blood Pressure (Mean BP)	969,062	77.9	15.3
Temperature, °F	209,499	98.5	1.4
Creatinine	67,565	1.5	1.2
Blood Urea Nitrogen (BUN)	66,746	31.0	21.1
White Blood Cell Count (WBC)	59,777	11.6	6.2
Lactate	39,667	2.4	1.8

Included in the 20 physiological traits we filter for are eight that are particularly predictive of the onset of severe sepsis, septic shock, or acute kidney failure. These traits are included in the SIRS (System Inflammatory Response Syndrome) and SOFA (Sequential Organ Failure Assessment) criteria [17]. The average daily measurements or lab test orders across the chosen cohort for these eight traits is highly variable (Figure 1). Of these eight traits, the first three are vitals measured using bedside monitoring systems for which approximately hourly measurements are recorded; the latter four are labs requiring phlebotomy and are typically measured just 2–3 times each day. We find the frequency of orders also varies across different labs, possibly due in part to differences in cost; for example, WBC (which is relatively inexpensive to test) is on average sampled slightly more often than lactate. In order to apply our proposed RL

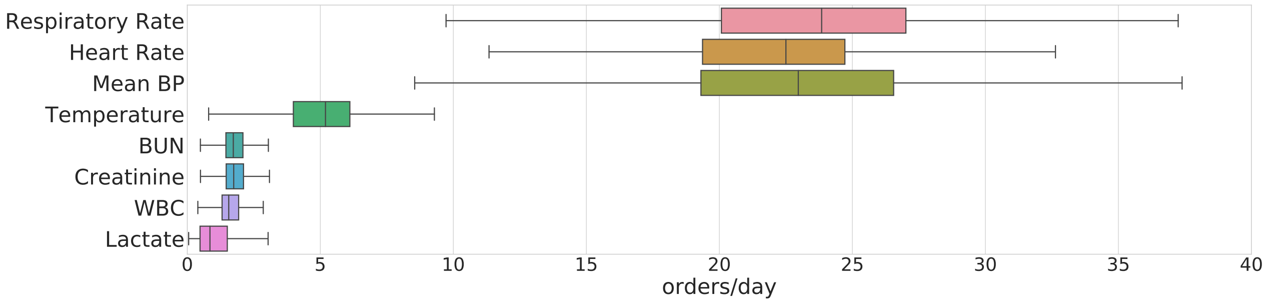

Fig. 1. **Mean number of recorded measurements per day, of chosen vitals and lab tests.** These eight traits are commonly used in computing clinical risk scores or diagnosing sepsis.

algorithm to this sparse, irregularly sampled dataset, we adapt the multi-output Gaussian process (MOGP) framework [10] to obtain hourly predictions of patient state with uncertainty quantified, on 17 of the 20 clinical traits. For three of the vitals, namely the components of the Glasgow Coma Scale, we impute with the last recorded measurement.

2.2. *MDP formulation*

Each patient admission is modelled as an MDP with:

(1) a state space $\mathcal{S}$, such that the patient physiological state at time t is given by $s_t \in \mathcal{S}$;
(2) an action space $\mathcal{A}$ from which the clinician's action a_t is chosen;
(3) an unknown transition function $\mathcal{P}_{sa}$ that determines the patient dynamics; and
(4) a reward function r_t that constitutes the observed clinical feedback for this action.

The objective of the RL agent is to learn an optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected discounted accumulated reward over the course of an admission:

$$\pi^* = \operatorname*{argmax}_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t | \pi\right], \text{ where } T \text{ is admission length, } \gamma \text{ is the discount factor.}$$

We start by describing the state space of our MDP for ordering lab tests. We first resample the raw time series using a multi-objective Gaussian process with a sampling period of one hour. The patient state at time t is defined by:

$$\boldsymbol{s_t} = \left[m_t^{SOFA}, \boldsymbol{m}_t^{vitals}, \boldsymbol{m}_t^{labs}, \boldsymbol{y}_t^{labs}, \boldsymbol{\Delta}_t^{labs}\right]^\top \tag{1}$$

Here, $\boldsymbol{m}_t$ denotes the predictive means and standard deviations respectively of each of the vitals and lab tests. For the predictive SOFA score m_t^{SOFA}, we compute the value using its clinical definition, from the predictive means on five traits—mean BP, bilirubin, platelet, creatinine, FiO_2—along with GCS and related medication history (e.g., dopamine). Vitals include any time-varying physiological traits that we consider when determining whether to order a lab test. Here, we look at four key physiological traits—heart rate, respiratory rate, temperature, and mean blood pressure—and four lab tests—creatinine, BUN, WBC, and lactate. The values $\boldsymbol{y}_t$ are the last known measurements of each of the four labs, and $\boldsymbol{\Delta}_t$ denotes the elapsed time since each was last ordered. This formulation results in a 21-dimensional state space. Depending on the labs that we wish to learn recommendations for testing, the action space $\mathcal{A}$

is a set of binary vectors whose 0/1 elements indicate whether or not to place an order for a specific lab. These actions can be written as $\boldsymbol{a_t} \in \mathcal{A} = \{1, 0\}^L$, where L is the number of labs. In our experiments, we learn policies for each of the four labs independently, such that $L = 1$, but this framework could be easily extended to jointly learning recommendations for multiple labs.

In order for our RL agent to learn a meaningful policy, we need to design a reward function that provides positive feedback for the ordering of tests where necessary, while penalizing the over- or under-ordering of any given lab test. In particular, the agent should be encouraged to order labs when the physiological state of the patient is abnormal with high probability, based on estimates from the MOGP, or when a lab is predicted to be informative (in that the forecasted value is significantly different from the last known measurement) due to a sudden change in disease state. In addition, the agent should incur some penalty whenever a lab test is taken, decaying with elapsed time since the last measurement, to reflect the effective cost (both economic and in terms of discomfort to the patient) of the test. We formulate these ideas into a vector-valued reward function $\boldsymbol{r_t} \in \mathbb{R}^d$ of the state and action at time t, as follows:

$$\boldsymbol{r_t} = \left[{r_t}^{SOFA}\ , {r_t}^{treat}\ , {r_t}^{info}\ , -{r_t}^{cost}\ \right]^\top \tag{2}$$

Patient state: The first element, r^{SOFA}, uses the recently introduced SOFA score for sepsis [18] which assesses severity of organ dysfunction in a potentially septic patient. Our use of SOFA is motivated by the fact that, in practice, sepsis is more often recognized from the associated organ failure than from direct detection of the infection itself [19]. The raw SOFA score ranges from 0 to 24, with a maximum of four points assigned each to symptom of failure in the respiratory system, nervous system, liver, kidneys, and blood coagulation. A change in SOFA score ≥ 2 is considered a critical index for sepsis [18]. We use this rule of thumb to design the first reward term as follows:

$${r_t}^{SOFA} = \mathbb{1}_{\boldsymbol{a_t} \neq \boldsymbol{0}} \cdot \mathbb{1}_{f(\cdot) \geq 2}\ , \text{ where } f(\cdot) = m_t^{SOFA} - m_{t-1}^{SOFA}. \tag{3}$$

The raw score m_t^{SOFA} at each time step t is evaluated using current patient labs and vitals [19].

Treatment onset: The second term is an indicator variable for rewards capturing whether or not there is some treatment or intervention initiated at the next time step, $\boldsymbol{s_{t+1}}$:

$${r_t}^{treat} = \mathbb{1}_{\boldsymbol{a_t} \neq \boldsymbol{0}} \cdot \sum_{i \in M} \mathbb{1}_{\boldsymbol{s_{t+1}}(\text{treatment } i \text{ was given})}, \tag{4}$$

where M denotes the set of disease-specific interventions of interest. Again, the reward term is positive if a lab is ordered; this is based on the rationale that, if a lab test is ordered and immediately followed by an intervention, the test is likely to have provided actionable information. Possible interventions include antibiotics, vasopressors, dialysis or ventilation.

Lab redundancy: The term ${r_t}^{info}$ denotes the feedback from taking one or more lab tests with novel information. We quantify this by using the mean absolute difference between the last observed value and predictive mean from the MOGP as a proxy for the information available:

$${r_t}^{info} = \sum_{\ell=1}^{L} max\left(0, g(\cdot) - c_\ell\right) \cdot \mathbb{1}_{\boldsymbol{a_t}[\ell]=1}\ , \text{ where } g(\cdot) = \left| \frac{m_t^{(\ell)} - y_t^{(\ell)}}{\sigma_t^{(\ell)}} \right|, \tag{5}$$

where σ_t^ℓ is the normalization coefficient for lab ℓ, and the parameter c_ℓ determines the minimum prediction error necessary to trigger a reward; in our experiments, this is set to the median prediction error for labs ordered in the training data. The larger the deviation from current forecasts, the higher the potential information gain, and in turn the reward if the lab is taken.

Lab cost: The last term in the reward function, ${r_t}^{cost}$ adds a penalty whenever any test is ordered to reflect the effective "cost" of taking the lab at time t.

$$ {r_t}^{cost} = \sum_{\ell=1}^{L} \exp\left(-\frac{\Delta_t^{(\ell)}}{\Gamma_\ell}\right) \cdot \mathbb{1}_{\boldsymbol{a_t}[\ell]=1}, \tag{6} $$

where Γ_ℓ is a decay factor that controls the how fast the cost decays with the time Δ_t elapsed since the last measurement. In our experiments, we set $\Gamma_\ell = 6 \; \forall \ell \in L$.

2.3. *Learning optimal policies*

Once we extract sequences of states, actions, and rewards from the ICU data, we can generate a dataset of one-step transition tuples of the form $\mathcal{D} = \{\langle s_t^n, a_t^n, s_{t+1}^n \rangle, r_t^n\}$, $n = 1...|\mathcal{D}|$. These tuples can then be used to learn an estimate of the Q-function, $\hat{Q} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$ —where $d = 4$ is the dimensionality of the reward function—to map a given state-action pair to a vector of expected cumulative rewards. Each element in the Q-vector represents the estimated value of that state-action pair according to a different objective. We learn this Q-function using a variant of Fitted Q-iteration (FQI) with extremely randomized trees [13, 20]. FQI is a batch off-policy reinforcement learning algorithm that is well-suited to clinical applications where we have limited data and challenging state dynamics. The algorithm adapted here to handle vector-valued rewards is based on Pareto-optimal Fitted-Q [15].

In order to scale from the two-stage decision problem originally tackled to the much longer admission sequences here (≥ 24 time steps), we define a stricter pruning of actions: at each iteration we eliminate any *dominated* actions for a given state—those actions that are outperformed by alternatives for all elements of the Q-function—and retain only the set $\Pi(s) = \{a : \nexists a' \, (\forall \, d, \; \hat{Q}_d(s,a) < \hat{Q}_d(s,a'))\}$ for each s. Actions are further filtered for *consistency*: we might consider feature consistency to be defined as rewards being linear in each feature space [15]. Here, we relax this idea to filter out only those actions from policies that cannot be expressed by our nonlinear tree-based classifier. The function will still yield a non-deterministic policy (NDP) as, in most cases, there will not be a strictly optimal action that achieves the highest Q_d for all d. We suggest one possible approach for reducing the NDP to give a single best action for any given state based on practical considerations in the next section.

3. Results

Following the extraction of our 6,060 admissions and resampling in hourly intervals using the forecasting MOGP, we partitioned the cohort into training and test sets of 3,636 and 2,424 admissions respectively. This gave approximately 500,000 one-step transition tuples of the form $\langle s_t, a_t, s_{t+1}, r_t \rangle$ in the training set, and over 350,000 in the test set. We then ran batched FQI with these samples for 200 iterations with discount factor $\gamma = 0.9$. Each iteration took

Algorithm 1 Multi-Objective Fitted Q-iteration with strict pruning (MO-FQI)

```
Input:
One-step transitions $\mathcal{F} = \{\langle s_t^n, a_t^n, s_{t+1}^n \rangle, r_{t+1}^n\}_{n=1:|\mathcal{F}|}$;
Regression parameters $\theta$; action space $\mathcal{A}$; subset size $N$
Initialize $Q^{(0)}(s_t, a_t) = \mathbf{0} \in \mathbb{R}^d \quad \forall s_t \in \mathcal{F},\ a_t \in \mathcal{A}$
for iteration $k = 1 \rightarrow K$ do
    Sample $subset_N \sim \mathcal{F}$ ; initialize $S \leftarrow []$
    for $i \in subset_N$ do
        Generate set $\Pi(s_i)$ using $Q^{(k-1)}$
        Initialize classification parameters $\phi$
        $\phi \leftarrow classify(s_i, a_i)$
        for $\pi_i \in \Pi$ : do
            $a' \leftarrow \pi_i(s_{i+1}) \cap predict(s_{i+1}, \phi)$
            $Q^{(k)}(s_i, a_i) \leftarrow r_{i+1} + \gamma Q^{(k-1)}(s_{i+1}, a')$
        end
        $S \leftarrow append(S, \langle (s_i, a_i),\ Q^{(k)}(s_i, a_i) \rangle)$
    end
    $\theta \leftarrow regress(S)$
end
Result: $\theta$
```

100,000 transitions, sampled from the training set, with probability inversely proportional to the frequency of the action in the tuple. The vector-valued outputs of estimated Q-function were then used to obtain a non-deterministic policy for each lab considered (Section 2.3). We chose to collapse this set to a practical deterministic policy as follows:

$$\Pi(s) = \begin{cases} 1, & \hat{Q}_d(s, a=0) < \hat{Q}_d(s, a=1) + \varepsilon_d, \quad \forall\, d \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

In particular, a lab should be taken ($\Pi(s) = 1$) *only if* the action is *optimal*, or estimated to outperform the alternative for all objectives in the Q-function. This strong condition for ordering a lab is motivated by the fact that one of our primary objectives here is to minimize unnecessary ordering; the variable ε_d allows us to relax this for certain objectives if desired. For example, if cost is a softer constraint, setting $\varepsilon_{cost} > 0$ is an intuitive way to specify this preference in the policy. In our experiments, we tuned ε_{cost} such that the total number of recommended orders of each lab approximates the number of actual orders in the training set.

With a deterministic set of optimal actions, we could train our final policy function $\pi : \mathcal{S} \rightarrow A$; again, we used extremely randomized trees. The estimated Gini feature importances of the policies learnt show that in the case of lactate the most important features are the mean and measured lactate, the time since last lactate measurement (Δ) and the SOFA score (Figure 2). These relative importance scores are expected: a change in SOFA score may indicate the onset of sepsis, and in turn warrant a lactate test to confirm a source of infection, fitting typical clinical protocol. For the other three policies (WBC, creatinine, BUN) again the time since last measurement of the respective lab tends be prominent in the policy, along with the Δ terms

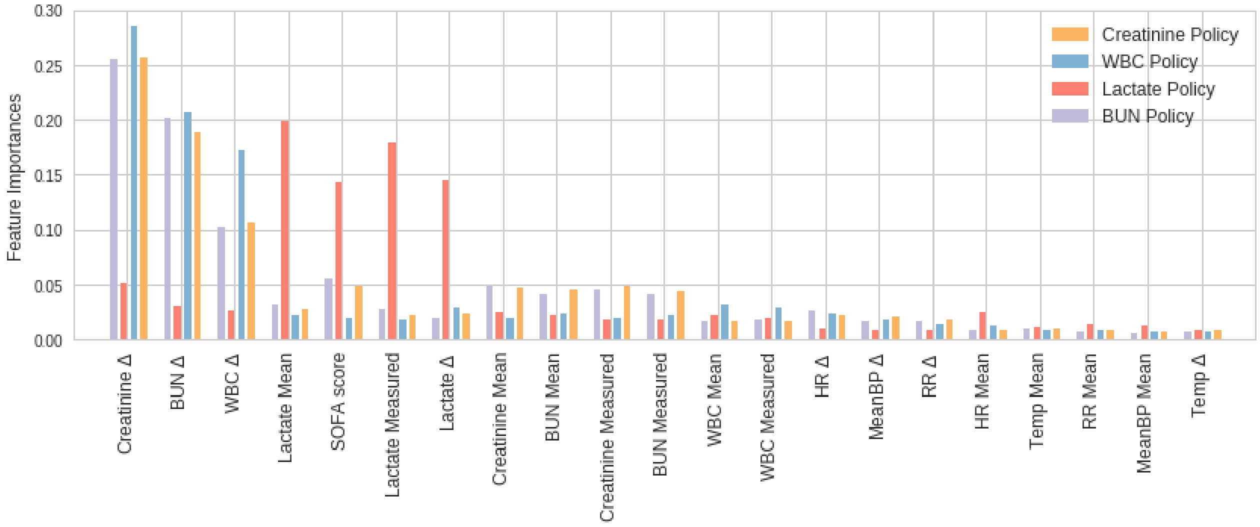

Fig. 2. **Feature importances** over the 21-dimensional state space, for each of our four policies.

for the other two labs. This suggests an overlap in information in these tests: For example, abnormally high white blood cell count is a key criteria for sepsis; severe sepsis often cascades into renal failure, which is typically diagnosed by elevated BUN and creatinine levels [21].

Once we have trained our policy functions, an additional component is added to our final recommendations: we introduce a *budget* that suggests taking a lab at the end of every 24 hour period for which our policy recommends no orders. This allows us to handle regions of very sparse recommendations by the policy function, and reflects clinical protocols that require minimum daily monitoring of key labs. In the policy for lactate orders in a typical patient admission, looking at the timing of the actual clinician orders, recommendations from our policy, and suggested orders from the budget framework, the actions are concentrated where lactate values are increasingly abnormal, or at sharp rises in SOFA score (Figure 3).

3.1. *Off-Policy Evaluation*

We evaluated the quality of our final policy recommendations in a number of ways. First, we implemented the per-step weighted importance sampling (PS-WIS) estimator to calculate the value of the policy π_e to be evaluated:

$$\hat{V}_{\text{PS-WIS}}(\pi_e) = \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma_{\text{WIS}}^{t} \left[\frac{\rho_t^{(i)}}{\sum_{i=1}^{n} \rho_t^{(i)}} \right] r_t^{(i)}, \quad \text{where } \rho_t = \prod_{j=0}^{t-1} \frac{\pi_e(s_j|a_j)}{\pi_b(s_j|a_j)},$$

given data collected from behaviour policy π_b [22]. The behaviour policy was found by training a regressor on real state-action pairs observed in the dataset. The discount factor was set to $\gamma_{\text{WIS}} = 1.0$, so all time steps contribute equally to the value of a trajectory.

We then compared estimates for our policy (MO-FQI) against the behaviour policy and a set of randomized policies as baselines. These randomized policies were designed to generate random decisions to order a lab, with probabilities $p = \{0.01, p_{emp}, 0.5\}$, where p_{emp} is the empirical probability of an order in the behaviour policy. For each p, we evaluated ten randomly generated policies and averaged performance over these. We observed that MO-FQI outperforms

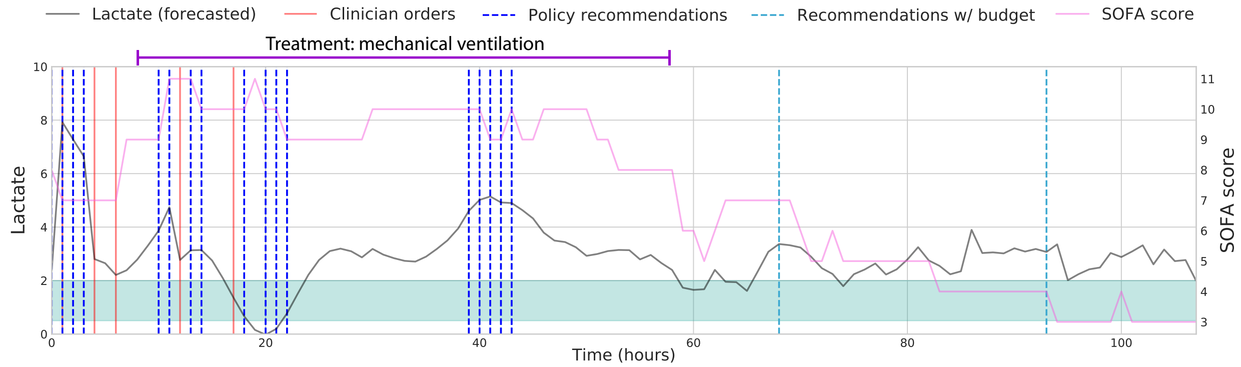

Fig. 3. **Demonstration of one test trajectory of recommending lactate orders.** The shaded green region denotes the range of normal lactate values (0.5–2 mmol/L).

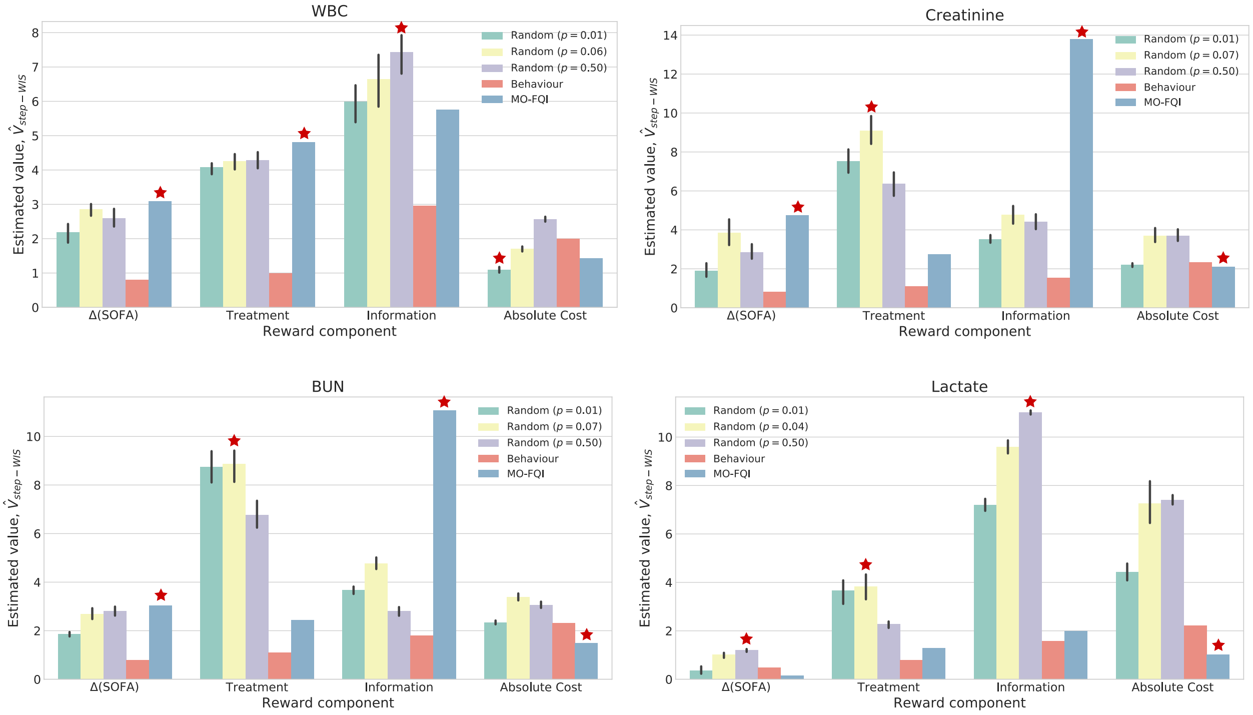

Fig. 4. **Evaluating $\hat{V}_d(\pi_e)$** for each reward component d, across policies for four labs. For randomized policies, error bars show standard deviations across 10 trials. The ($\star$) indicates the best performing policy for each reward component; for *absolute* cost, this corresponds to the lowest estimated value.

the behaviour policy across all reward components, for all four labs (Figure 4). Our policy also consistently approximately matches or outperforms other policies in terms of cost—note that lower cost is better—even with the inclusion of the slack variable ε_{cost} and the budget framework. Across the remaining objectives, MO-FQI outperforms the random policy in at least two of three components for all but lactate. This may be due in part to the relatively sparse orders for lactate resulting in higher variance value estimates.

In addition to evaluating using the per-step WIS estimator, we looked for more intuitive measures of how the final policy influences clinical practice. We computed three metrics here: (i) estimated reduction in total number of orders, (ii) mean information gain of orders taken, and (iii) time intervals between labs and subsequent treatment onsets.

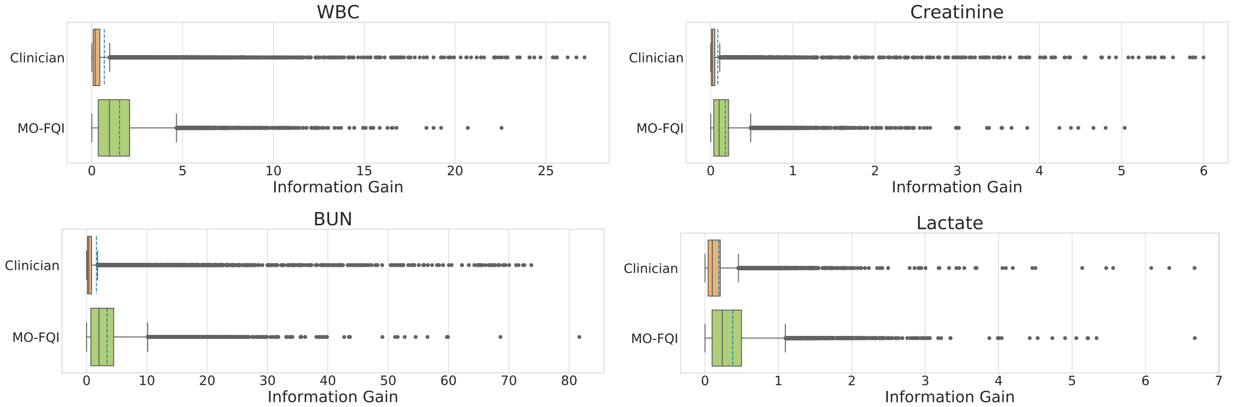

Fig. 5. **Evaluating Information Gain** of clinician actions against MO-FQI across all labs: the mean information in labs ordered by clinicians is consistently outperformed by MO-FQI: 0.69 vs 1.53 for WBC; 0.09 vs 0.18 for creatinine; 1.63 vs 3.39 for BUN; 0.19 vs 0.38 for lactate.

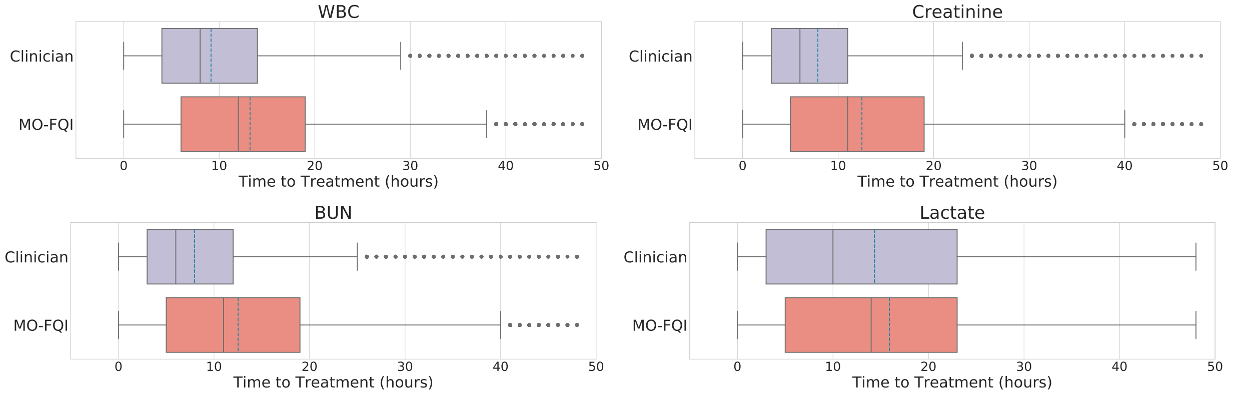

Fig. 6. **Evaluating Time to Treatment Onset** of lab orders by the clinician against MO-FQI across all labs: the mean time intervals are as follows (Clinician vs MO-FQI): 9.1 vs 13.2 for WBC; 7.9 vs 12.5 for creatinine; 8.0 vs 12.5 for BUN; 14.4 vs 15.9 for lactate.

In evaluating the total number of recommended orders, we first filter a sequence of recommended orders to the just the first (onset) of recommendations if there are no clinician orders between them. We argue that this is a fair comparison as subsequent recommendations are made without counterfactual state estimation, i.e., without assuming that the first recommendation was followed the clinician. Empirically, we find that the total number of recommendations is considerably reduced. For instance, in the case of recommending WBC orders, our final policy reports 12,358 orders in the test set, achieving a reduction of 44% from the number of true orders (22,172). In the case of lactate, for which clinicians' orders are the least frequent (14,558), we still achieved a reduction of 27%.

We also compared the approximate information gain of the actions taken by the estimated policy, in comparison with the policy used in the collected data. To do this, we defined the information gain at a given time by looking at the difference between the *approximated* true value of the target lab, which we impute using the MOGP model given all the observed values, and the forecasted value, computed using only the values observed before the current time. The distribution of aggregate information gain for orders recommended by our policy and actual clinician's orders in the test set shows higher mean information gain with MO-FQI (Figure 5).

Lastly, we considered the time to onset of critical interventions, which we define to include initiation of vasopressors, antibiotics, mechanical ventilation or dialysis. We first obtained a sequence of treatment onset times for each test patient; for each of these time points, we traced back to the earliest observed or recommended order taking place within the past 48 hours, and computed the time between these: $\Delta_t = t_{treatment} - t_{order}$. The distribution of time-to-treatment for labs taken by the clinician in the true trajectory against that for recommendations from our policy, for all four labs, shows that the recommended orders tend to happen earlier than the actual time of an order by the clinician—on average over an hour in advance for lactate, and more that four hours in advance for WBC, creatinine, and BUN (Figure 6).

4. Conclusion

In this work, we propose a reinforcement learning framework for decision support in the ICU that learns a compositional optimal treatment policy for the ordering of lab tests from sub-optimal histories. We do this by designing a multi-objective reward function that reflects clinical considerations when ordering labs, and adapting methods for multi-objective batch RL to learning extended sequences of Pareto-optimal actions. Our final policies are evaluated using importance-sampling based estimators for off-policy evaluation, metrics for improvements in cost, and reducing redundancy of orders. Our results suggest that there is considerable room for improvement on current ordering practices, and the framework introduced here can help recommend best practices and be used to evaluate deviations from these across care providers, driving us towards more efficient health care. Furthermore, the low risk of these types of interventions in patient health care reduces the barrier of testing and deploying clinician-in-the-loop machine learning-assisted patient care in ICU settings.

References

1. T. Badrick, Evidence-based laboratory medicine *The Clinical Biochemist Reviews* **34** (The Australian Association of Clinical Biochemists, 2013).
2. M. Zhi, E. L. Ding, J. Theisen-Toupal, J. Whelan and R. Arnaout, The landscape of inappropriate laboratory testing: a 15-year meta-analysis *PloS one* **8** (Public Library of Science, 2013).
3. T. Loftsgard and R. Kashyap, Clinicians role in reducing lab order frequency in icu settings *J Perioper Crit Intensive Care Nurs* **2**2016.
4. R. L. Konger, P. Ndekwe, G. Jones, R. P. Schmidt, M. Trey, E. J. Baty, D. Wilhite, I. A. Munshi, B. M. Sutter, M. Rao *et al.*, Reduction in unnecessary clinical laboratory testing through utilization management at a us government veterans affairs hospital *American journal of clinical pathology* **145** (Oxford University Press, 2016).
5. ICUMedical, Reducing the risk of iatrogenic anemia and catheter-related bloodstream infections using closed blood sampling (ICU Medical Inc., 2015).
6. J. Lee and D. M. Maslove, Using information theory to identify redundancy in common laboratory tests in the intensive care unit *BMC medical informatics and decision making* **15** (BioMed Central, 2015).

REFERENCES

7. F. Cismondi, L. A. Celi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa and S. N. Finkelstein, Reducing unnecessary lab testing in the icu with artificial intelligence *International journal of medical informatics* **82** (Elsevier, 2013).
8. Y. Luo, P. Szolovits, A. S. Dighe and J. M. Baron, Using machine learning to predict laboratory test results *American Journal of Clinical Pathology* **145**2016.
9. M. Ghassemi, M. A. F. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits and M. Feng, A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
10. L.-F. Cheng, G. Darnell, C. Chivers, M. Draugelis, K. Li and B. Engelhardt, Sparse multi-output gaussian processes for medical time series prediction (2017).
11. S. Nemati, M. M. Ghassemi and G. D. Clifford, Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach, in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, 2016.
12. A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits and M. Ghassemi, Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach, in *Proceedings of the Machine Learning for Health Care, MLHC 2017, Boston, Massachusetts, USA, 18-19 August 2017*, 2017.
13. N. Prasad, L. Cheng, C. Chivers, M. Draugelis and B. Engelhardt, A reinforcement learning approach to weaning of mechanical ventilation in intensive care units, in *Uncertainty in Artificial Intelligence 2017*, 1 2017.
14. S. Natarajan and P. Tadepalli, Dynamic preferences in multi-criteria reinforcement learning, in *Proceedings of the 22nd international conference on Machine learning*, 2005.
15. D. J. Lizotte and E. B. Laber, Multi-objective markov decision processes for data-driven decision support *Journal of Machine Learning Research* **17**2016.
16. A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G. Mark, Mimic-iii, a freely accessible critical care database *Scientific data* **3** (Nature Publishing Group, 2016).
17. P. E. Marik and A. M. Taeb, *Journal of thoracic disease* **9**, p. 943 (2017).
18. M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, The third international consensus definitions for sepsis and septic shock (sepsis-3) *Jama* **315** (American Medical Association, 2016).
19. J.-L. Vincent, G. S. Martin and M. M. Levy, qsofa does not replace sirs in the definition of sepsis *Critical Care* **20** (BioMed Central, 2016).
20. D. Ernst, P. Geurts and L. Wehenkel, Tree-based batch mode reinforcement learning *Journal of Machine Learning Research* **6**2005.
21. M. R. Clarkson, B. M. Brenner and C. Magee, *Pocket Companion to Brenner and Rector's The Kidney E-Book* (Elsevier Health Sciences, 2010).
22. D. Precup, R. S. Sutton and S. P. Singh, Eligibility traces for off-policy policy evaluation, in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000).