Contextualized Policy Recovery: Modeling and Interpreting Medical Decisions with Adaptive Imitation Learning

Jannik Deuschel [* 1 2]   **Caleb N. Ellington** [* 1]   **Yingtao Luo** [1]   **Benjamin J. Lengerich** [3 4]   **Pascal Friederich** [2]
Eric P. Xing [5 1 6]

Abstract

Interpretable policy learning seeks to estimate intelligible decision policies from observed actions; however, existing models force a tradeoff between accuracy and interpretability, limiting data-driven interpretations of human decision-making processes. Fundamentally, existing approaches are burdened by this tradeoff because they represent the underlying decision process as a universal policy, when in fact human decisions are dynamic and can change drastically under different contexts. Thus, we develop Contextualized Policy Recovery (CPR), which re-frames the problem of modeling complex decision processes as a multi-task learning problem, where each context poses a unique task and complex decision policies can be constructed piece-wise from many simple context-specific policies. CPR models each context-specific policy as a linear map, and generates new policy models *on-demand* as contexts are updated with new observations. We provide two flavors of the CPR framework: one focusing on exact local interpretability, and one retaining full global interpretability. We assess CPR through studies on simulated and real data, achieving state-of-the-art performance on predicting antibiotic prescription in intensive care units (+22% AUROC vs. previous SOTA) and predicting MRI prescription for Alzheimer's patients (+7.7% AUROC vs. previous SOTA). With this improvement, CPR closes the accuracy gap between interpretable and black-box methods, allowing high-resolution exploration and analysis of context-specific decision models.

---

[*]Equal contribution [1]Carnegie Mellon University [2]Karlsruhe Institute of Technology [3]Broad Institute of MIT and Harvard [4]MIT [5]MBZUAI [6]Petuum, Inc.. Correspondence to: Caleb N. Ellington <cellingt@cs.cmu.edu>, Eric P. Xing <epxing@cs.cmu.edu>.
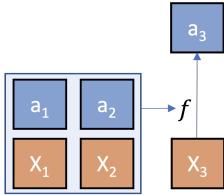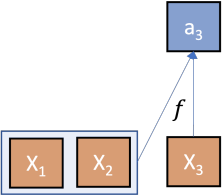
1. Introduction

Interpretable policy learning (Hüyük et al., 2022) seeks to recover an underlying decision-making process from a dataset of demonstrated behavior, and represent this process as an interpretable model that can be quantified, audited, and intuitively understood. This approach has gained considerable attention in the medical informatics community as a promising approach for improving standards of care by detecting bias, explaining sub-optimal outcomes (Lengerich et al., 2022), and quantifying regional (McKinlay et al., 2007) and institutional (Westert et al., 2018) differences. Classic machine learning algorithms for policy inference are based on inverse reinforcement learning (Ng & Russell, 2000) or imitation learning (Bain & Sammut, 1999; Piot et al., 2014; Ho & Ermon, 2016), and use black-box architectures such as recurrent neural networks. These approaches have been applied in various medical domains, most prominently oncology prognosis (Beck et al., 2011; Esteva et al., 2017). However, black-box methods have been met with skepticism from the medical community based on a lack of interpretability as well as an inability to identify catastrophic failure modes and generalization issues (Laï et al., 2020; Royal Society (Great Britain) and Royal Society (Great Britain) Staff, 2017).

To address this, several recent works have proposed highly constrained policy parametrizations for interpretability. These parameterizations take the form of recurrent decision trees (Pace et al., 2022), visual decision boundaries (Hüyük et al., 2022), high-level programming syntax (Verma et al., 2018) or outcome preferences (Yau et al., 2020). Highly constrained models are generally more interpretable, but their interpretability stems from restrictive modeling architectures, sacrificing performance or imposing obscure constraints that make real-world applications challenging. The primary challenge is that human decisions are informed by a variety of factors including patient background, medical history, lab tests, and more, and true human decision processes are complex. Thus, compressing the decision policy into a single universal observation-to-action mapping necessitates the use of large nonparametric models (e.g. neural nets) that preclude direct interpretability, or produce a model which

Policy class Representative model	Contextualized CPR	Global Tree POETREE	Global Aggregate INTERPOLE
$\mathbb{P}(a_t\mid x_{0:t}, a_{0:t-1})$	$f_{h(x_{0:t-1}, a_{0:t-1})}(x_t)$	$f_\theta(x_t, h(x_{0:t-1}))$	$f_\theta(h(x_{0:t}, a_{0:t-1}))$
Adapts to observed actions	✓	✗	✓
$x_t \to a_t$ glass-box	✓	✓	✗

Table 1. Comparison of interpretable policy learning algorithms. All listed model classes are designed to provide interpretability beyond what a black-box recurrent model would provide: CPR provides contextualized parametric policies, POETREE (Pace et al., 2022) provides a global tree-based policy, and INTERPOLE (Hüyük et al., 2022) provides interpretations of encoded belief states as a summary of patient history. While each model is considered loosely explainable, the adaptability of CPR's context-specific linear policies enable substantially simpler functional forms for $f(\cdot)$ than other methods, requiring no post-processing to interpret.

fails to capture the complexities of human decision-making. Succinctly, models must be both accurate *and* interpretable to effectively support clinical decisions.

In this paper, we propose Contextualized Policy Recovery (CPR). Instead of seeking a universal policy that necessitates trading off accuracy against interpretability, CPR embraces the wealth of contextual information that guides human decision-making and reframes the problem of policy learning as multi-task learning of interpretable *context-specific* policies. CPR learns a black-box generator function that encodes contextual information (e.g. the history of observed symptoms and actions which have previously occurred in the decision process) and generates linear observation-to-action mappings. This parameterization can be thought of as a hypernetwork (Ha et al.) for interpretable policy models. With this combination of black-box and glass-box components, CPR provides interpretable context-specific decision functions at sample-specific resolution. CPR does not sacrifice representational capacity and achieves state-of-the-art performance in policy recovery. Finally, CPR is a modular framework: both the model family of the context encoder and the interpretable observation-to-action mapping model can be chosen according to the given task at hand.

Contributions: Our work makes the following contributions to the personalized modeling and medical machine learning communities:

- We propose CPR, a framework for estimating time-varying and context-dependent policies as linear observation-to-action mappings, operating in a fully offline and partially observable environment. CPR enables interpretable and personalized imitation learning, dynamically incorporating new information to model decisions over the entire course of treatment.

- We apply CPR to two canonical medical imitation learning tasks: predicting antibiotic prescription in intensive care units and predicting MRI prescription for Alzheimer's patients. CPR matches the performance of black-box models, generating decision models that recover best-practice treatment policies under a continuum of patient contexts.

- We simulate a heterogeneous Markov decision process, where CPR empirically converges to the true decision model parameters, improving representation, performance, and interpretability over black-box models.

2. Related work

Expanding on the objective of classical imitation learning, which solely seeks to replicate demonstrated behavior in a partially observable and offline environment, we seek to learn an interpretable parametrization of observed behavior over time to understand why an agent's action was taken. CPR combines the strengths of previous work (Table 1) by keeping the observation-to-action mapping at each timestep interpretable while being able to adapt to the full observed past, aligning learned policies closer with demonstrated behavior.

Imitation Learning   The classical approach for learning sequential decision-making processes involves reinforcement learning, optimizing an agent's reward signal R in an online environment. However, in some applications, such as clinical decision-making, experiments with online policies would be both unethical and impractical so only observational data is available. In this setting we do not have access

to the reward signal R and instead focus on the inverse task of replicating the observed behavior of an agent, known as imitation learning. There are several approaches to tackle such problems, including simple behavioral cloning, without taking interpretability into account, in which the task is reduced to supervised learning mapping observations to actions (Bain & Sammut, 1999; Piot et al., 2014; Sun et al., 2017). Other approaches are based on distribution matching where adversarial training is used to match the state-action distributions between demonstrator and learned policies (Ho & Ermon, 2016; Jeon et al., 2018; Kostrikov et al., 2020). Adaptations of inverse reinforcement learning have been proposed for partial observability (Choi & Kim, 2011) and offline learning (Makino & Takeuchi, 2012). These approaches primarily utilize black-box models to maximize action-matching performance, making it challenging to distill a transparent and tractable description of the learned policies.

Interpretable Policy Learning Recurrent neural networks depend on latent hidden representations for each feed-forward pass in a time series, obfuscating popular post-hoc explanation methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). Instead, recent interpretable policy learning methods have focused on adding limiting assumptions, although this also limits real-world applicability. INTERPOLE (Hüyük et al., 2022) is a notable example, which parameterizes a latent belief space that relates to decisions, but falls short of explaining how this belief relates to prior observed information. Another notable method is POETREE (Pace et al., 2022), which parameterizes policies as a recurrent soft decision trees (Frosst & Hinton, 2017). While this approach results in an interpretable model which maps between timeseries observations and actions, POETREE requires significant post-processing to remove uninterpretable components critical to training. This post-processing step sacrifices performance, especially if the observation space is high-dimensional. Further, these methods are often less accurate at decision modeling than workhorse models like logistic regression (Table 1).

3. Methods

Prior imitation learning methods represent decision processes as a single universal policy. Accurate models of this policy are too complex to interpret, and interpretable models are too simple to be accurate. The difficulty in representing this policy relates to the dynamic nature of human decisions, which can change drastically under different contexts. However, considering a single decision *in context* is relatively simple. Given a patient's clinical profile, medical history, known risk factors, and prior treatments, there are only a few reasonable diagnoses and possible treatments. By learning to predict patient-specific policies based on each

patient's context, we can substantially simplify the representation of the policy for each patient. Hence, we transform the problem of representing a complex universal policy into a problem of learning a simple policy model for every possible context. To achieve this, we apply contextualized modeling (Lengerich et al., a), which allows us to synthesize components from previous accurate and interpretable modeling methods in a single framework to learn interpretable context-specific policies.

3.1. Preliminaries: Contextualized Modeling

Due to an increase in dataset complexity, heterogeneity, and size, context-specific inference has driven interest in many application areas (Ageenko et al., 2010; Buettner et al., 2015; Fisher et al., 2018; Hart, 2016; Ng et al., 2015). Given a dataset consisting of targets $y \in Y$, predictors $x \in X$ and context $c \in C$, with the corresponding random variables denoted as $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{C}$, we want to learn a model $\mathbb{P}(\mathbf{Y} \mid x, c)$ predicting y from x and c. Contextualized modeling assumes a hierarchical model

$$y \sim \mathbb{P}(\mathbf{Y} \mid x, \theta), \quad \theta \sim \mathbb{P}(\theta \mid c)$$
$$\mathbb{P}(\mathbf{Y} \mid x, c) = \int_\theta \mathbb{P}(\mathbf{Y} \mid x, \theta) \mathbb{P}(\theta \mid c) d\theta$$

where θ defines the parameters of a context-specific model $P(\mathbf{Y} \mid x, \theta)$. This formulation was initially proposed by (Al-Shedivat et al., 2020) for context-specific linear models, and recently generalized to any model by (Lengerich et al., a). Contextualization allows us to use a simple parametric model class to represent the context-specific model $\mathbb{P}(\mathbf{Y} \mid x, \theta)$, while also retaining representational capacity across contexts by implementing $\mathbb{P}(\theta \mid c)$ as a rich model class, such as neural networks. For simplicity, $\mathbb{P}(\theta \mid C)$ is typically implemented as a Diroc delta function $\delta(\theta - g(C))$, where $g(C)$ is a deterministic context encoding. Many well-known contextualized models fit this framework. In varying coefficient models (Hastie & Tibshirani, 1993) g is a linear model or a kernel method. In contextual explanation networks (Al-Shedivat et al., 2020; Lengerich et al., b) g is a deep neural network and $\mathbb{P}(\mathbf{Y}|x, \theta)$ is linear. In hypernetworks (Ha et al.) $\mathbb{P}(\mathbf{Y}|x, \theta)$ is also a neural network. In the contextual GAM (Lengerich et al., c), g is a neural additive model (Agarwal et al.). In contextualized graphical models (Ellington et al.; Lengerich et al., 2021), $\mathbb{P}(\mathbf{Y}, \mathbf{X} \mid \theta)$ is a joint distribution on $\mathbf{Y}$ and $\mathbf{X}$. In CPR, g is a differentiable, recurrent model, such as a recurrent neural network, and $\mathbb{P}(\mathbf{Y}|x, \theta)$ is a logistic regression model.

3.2. Contextualized Policy Recovery

CPR builds on recent developments in interpretable, offline policy learning (Pace et al., 2022). Let dataset $D = \{(x_1^i, a_1^i), ..., (x_{T_i}^i, a_{T_i}^i)\}_{i=1}^N$ consist of N treatment trajectories, where each patient i is observed over T_i discrete

timesteps for symptoms $x \in \mathcal{X}$ and physician actions $a \in \mathcal{A}$. The data is generated by an unknown policy of the physician agent $\mathbb{P}(a_t | x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t)$ where the action probability at time t is a function of the agent's current state, which is defined by the current and past patient symptoms, as well as past actions.

To recover a policy that is both tractable and interpretable, CPR makes a practical enabling assumption: in many real world settings, recent information has the highest importance when making a decision, and historical information contextualizes the effect of new information. In medicine, treatment history is useful to place a patient's current disease presentation within a context of disease progression and past treatment attempts, but decisions are always based on the current presentation. To represent this information hierarchy, CPR leverages contextual and historical features to generate context-specific policy models.

$$P(a_t | x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t) := f_{\theta_t}(a_t | x_t)$$
$$\theta_t = g(x_1, a_1, ..., x_{t-1}, a_{t-1})$$

Where f is an interpretable model class, e.g. logistic regression, parameterized by a context-specific θ, and θ is generated via a historical context encoder g. The effects of current observation x_t on action probabilities a_t are directly explained through the simple context-specific model f_θ. Furthermore, g can take any functional form without precluding the interpretability of f. The context-specific policy models $f_{\theta_t^i}$ are generated for each patient i at each timepoint t, allowing us to investigate how previous actions, previous symptoms, patient covariates, and treatment time influence the policy. CPR flexibly allows the context encoder g and the observation-to-action function f to be freely chosen, although they must be differentiable to allow for joint optimization under an appropriate loss ℓ.

$$\min_g \frac{1}{N} \sum_i \frac{1}{T_i} \sum_t \frac{1}{|\mathcal{A}|} \sum_{\widehat{a} \in \mathcal{A}} \ell(a_t^i, f_{g(x_1^i, a_1^i, ..., x_{t-1}^i, a_{t-1}^i)}(\widehat{a}, x_t^i))$$

In our experiments, g is parametrized by either a vanilla RNN or LSTM (Hochreiter & Schmidhuber, 1997), f is a logistic regression with coefficients θ, actions $\mathcal{A} := \{0, 1\}$ are binary, and ℓ is binary cross-entropy loss. While it is possible for the context-specific model f to be non-linear in reality, and while CPR also permits implementing any differentiable function for f, we implement f as a linear model based on the canonical nature of additive and linear approximations in both treatment effect estimation (Kreif et al.; Chernozhukov & Hansen; Ogburn et al.; Heckman et al., a;b) and explainability (Lundberg & Lee, 2017; Shrikumar et al., b;a; Ribeiro et al., 2016; Al-Shedivat et al., 2020). CPR is highly related to both areas, as we seek not only a linear explanation but the effect size of an observation in terms of its log-odds contribution toward an action. Finally, CPR applies a lasso regularizer to θ to learn robust policy parameters.

3.3. Global Interpretability

CPR synergizes black-box components with interpretable models to enable highly accurate policy models that are directly interpretable within a given context. While context-specific policies are a novel, accurate, and highly adaptable form of interpretability, there are cases where global interpretability is still necessary. To understand the exact effect of every historical feature on every action, we develop a second version of CPR which we call CPR global. In CPR global, we leverage the CPR framework to make piece-wise updates to a globally interpretable policy, which can be decomposed as a linear combination of all observed features and actions. Using the logistic policy form,

$$\text{logodds}(a_t | x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t) := \langle \theta_t, x_t \rangle + \mu_t$$
$$\theta_t = g(x_1, a_1, ..., x_{t-1}, a_{t-1})$$
$$\mu_t = \alpha \langle \beta_{t-1}, [x_{t-1}, a_{t-1}] \rangle + (1 - \alpha)\mu_{t-1}$$
$$\beta_{t-1} = h(x_1, a_1, ..., x_{t-2})$$

Where $\langle \cdot \rangle$ is the dot product and h is another recurrent context encoder. The bias term μ_t is updated with every new observation through a separate linear combination with β. The hyperparameter α weights this update against previous updates to the bias term. This form telescopes into a globally interpretable linear combination that is extended with each new observation. As a result, each observation has an exact linear effect in the context-specific global policy.

$$P(a_t | x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t) :=$$
$$\langle \theta_t, x_t \rangle + \sum_{k=0}^{t-1} \alpha(1 - \alpha)^k \langle \beta_t, [x_{t-k-1}, a_{t-k-1}] \rangle$$

4. Experiments

We apply CPR to recover time-varying, context-specific decision models within complex decision-making processes. First, we evaluate CPR on real MRI prescription data for dementia patients and antibiotic prescription data in intensive care units. Follow-up analysis of the contextualized models reveals best-practice treatment plans for both common and outlier patients, while also recovering unexpected and meaningful heterogeneity in physician policies. We further ensure CPR's ability to recover true policy models through a simulated heterogeneous Markov decision process.

4.1. Medical Datasets

We apply CPR to two medical datasets for canonical imitation learning tasks, ADNI and MIMIC-III. These datasets are a prime example of partially-observable decision environments in which we are forced to learn from demonstrated behavior, and where learned policies have the potential to improve clinical operations. CPR significantly outperforms

Algorithm	ADNI MRI scans			MIMIC Antibiotics		
	AUROC	AUPRC	Brier ↓	AUROC	AUPRC	Brier ↓
□ Logistic regression	0.66 ± 0.01	0.86 ± 0.00	0.16 ± 0.00	0.57 ± 0.01	0.80 ± 0.01	0.20 ± 0.00
□ INTERPOLE †	0.60 ± 0.04	0.81 ± 0.08	0.17 ± 0.05	NR	NR	NR
□ INTERPOLE ‡	0.44 ± 0.04	0.75 ± 0.09	0.19 ± 0.07	$0.65 \pm (\leq 0.04)$	NR	$0.21 \pm (\leq 0.04)$
□ POETREE ‡	0.62 ± 0.01	0.82 ± 0.01	0.18 ± 0.01	$0.68 \pm (\leq 0.04)$	NR	$0.19 \pm (\leq 0.04)$
□ CPR-RNN (ours)	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.90 \pm 0.00}$	$\mathbf{0.14 \pm 0.00}$
□ CPR-LSTM (ours)	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.90 \pm 0.00}$	$\mathbf{0.14 \pm 0.00}$
□ CPR-RNN global (ours)	$\mathbf{0.72 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.15 \pm 0.01}$	0.80 ± 0.01	0.89 ± 0.01	0.16 ± 0.00
□ CPR-LSTM global (ours)	$\mathbf{0.72 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.15 \pm 0.01}$	0.79 ± 0.01	0.89 ± 0.01	0.16 ± 0.00
■ RNN	0.72 ± 0.01	0.88 ± 0.01	0.15 ± 0.00	0.83 ± 0.00	0.90 ± 0.00	0.13 ± 0.00
■ LSTM	0.71 ± 0.01	0.88 ± 0.01	0.15 ± 0.00	0.84 ± 0.00	0.91 ± 0.00	0.13 ± 0.00

Table 2. Action-matching performance of imitation learning algorithms. Bolded values denote the best performance of interpretable models. Open source task-agnostic models are reported as mean ± standard error for 10 bootstrap runs. INTERPOLE and POETREE, which are task-specific or closed-source, are based on prior reports: † reported by (Hüyük et al., 2022), ‡ reported by (Pace et al., 2022). NR: No values reported. □: interpretable method. ■: black-box method.

other interpretable baseline models, and even performing on-par with fully black-box models for both EHR datasets (Table 2). Low brier scores indicate that CPR is well calibrated while achieving SOTA AUROC and AUPRC.

4.1.1. MIMIC ANTIBIOTICS

We look at 4195 patients in the intensive care unit over up to 6 timesteps extracted from the Medical Information Mart for Intensive Care III (Johnson et al., 2016) dataset and predict antibiotic prescription based on 7 observations - temperature, hematocrit, potassium, white blood cell count (WBC), blood pressure, heart rate, and creatinine. We removed the hemoglobin feature used in previous work by (Pace et al., 2022) since it is highly correlated (>0.95) with the hematocrit feature.

Contextualized policies reveal and explain heterogeneity in medical decision processes To see how decision functions change under different contexts, we compare them in *model space*. UMAP embeddings of the coefficient vectors of our context-specific policy models (Figure 1b) reveal three distinct clusters of decision functions. The rightmost cluster contains the initial model parameters θ_0 for each trajectory. Since there is no context that could differentiate the agent's behavior at the initial visit, these parameters are the same for all patients, and the contextualized models recover the population estimator. Subsequent to this initial homogeneity, heterogeneity in decision policies arises. The larger driver of this heterogeneity is prior antibiotic prescription – patients that previously got antibiotics are more likely to continue to receive antibiotics, while patients that did not receive antibiotics are likely to continue to not receive antibiotics. The lower cluster contains mostly (99.8%) models in which the patient did get antibiotics in the previous state $t-1$, while the upper cluster contains patients (99.3%) that

did not get antibiotics in $t-1$. This strong split is only recovered by contextualized policies; global policies that ignore context fail to identify this heterogeneity (Figure 1a). We train two models conditioned on their respective contexts, removing even more variability by limiting observations to the second day in the ICU. The global model only represents a small part of the population. Conditioning on the main driver of model heterogeneity (whether or not a patient got antibiotics in the previous visit) and training individual models for each case yields models that look like an average over the contextualized models of the respective clusters.

To uncover typical treatment regimes, we cluster patients into 5 subgroups over the first 4 days of the ICU stay using hierarchical clustering (Fig. 10). To identify the drivers of this heterogeneity in decision policies, we examine the parameters of the decision function as a function of context (Figure 1c). The most notable difference in parameters is the intercept value which is positive for the group of patients that get treated with antibiotics and negative for patients that did not receive antibiotic treatment. This is consistent across timesteps, and it is likely that a patient is prescribed antibiotics if they got it the day before. This aligns with medical protocols that rarely suggest antibiotic treatments briefer 5 days (Guleria et al., 2019).

Previous work by (Pace et al., 2022) and (Bica et al., 2021) described a patient's temperature and white blood cell count (WBC) as the main drivers of antibiotic treatment decisions since these are known medical criteria to counter infections (Masterton et al., 2008). Our results paint a more nuanced picture, in which heterogeneous coefficients reflect heterogeneous priorities when designing treatment plans. CPR identifies that the influence of temperature changes based on prior antibiotic prescription. For patients who have already been prescribed antibiotics, an infection has already been detected and the doctor's decision model shifts towards mit-

Contextualized Policies are Patient-specific and Capture Treatment Heterogeneity...
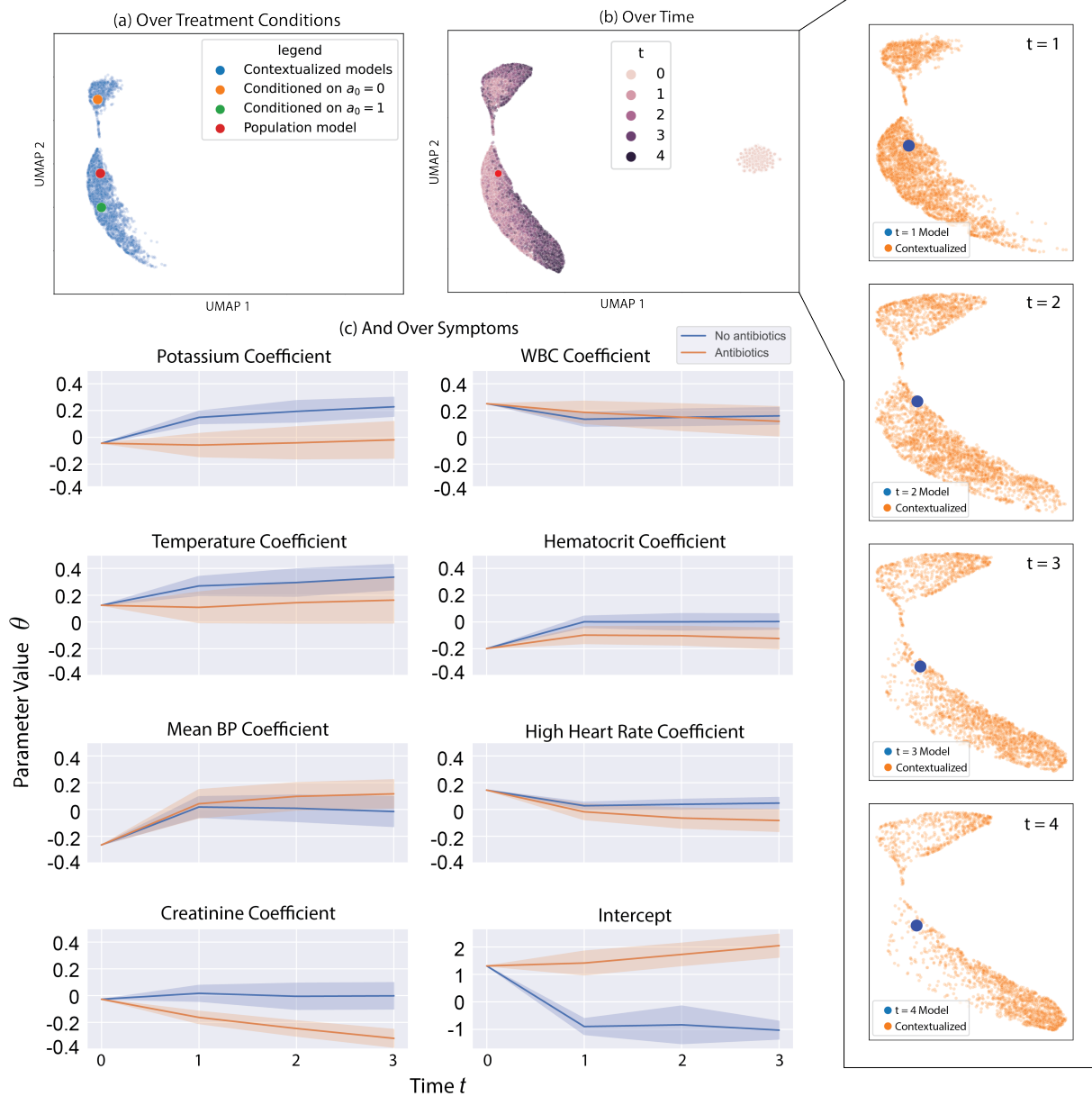

Figure 1. Exploration of contextualized policies generated by CPR for predicting antibiotic prescription. (a) Contextualized policies identify prior antibiotic prescription and (b) treatment time as drivers of treatment heterogeneity. We compare CPR's predicted policies at each timestep against population (left) and timestep-specific (right) logistic regression policies. (c) CPR generates policies that evolve with time and treatment history, revealing the context-specific importance of patient symptoms toward future treatments.

igating the risk of possible side effects of the antibiotics treatment rather than strictly considering the benefits. This shift in priorities is supported by the change in the creatinine coefficient (Figure 1c). High serum creatinine can be an indicator of impaired kidney function (Gounden et al., 2023), a possible adverse effect of antibiotics (Khalili et al., 2013); as such, a high creatinine level decreases the probability of continuing antibiotic treatment. Finally, we see

that increased potassium is associated with the decision to begin antibiotics (Figure 1c). Electrolyte balance (in which potassium plays a vital role) is an important factor in infectious diseases, and reduced sodium (a 1+ ion competitor with potassium) in particular is known as a marker of viral and bacterial infections (Królicka et al., 2020).
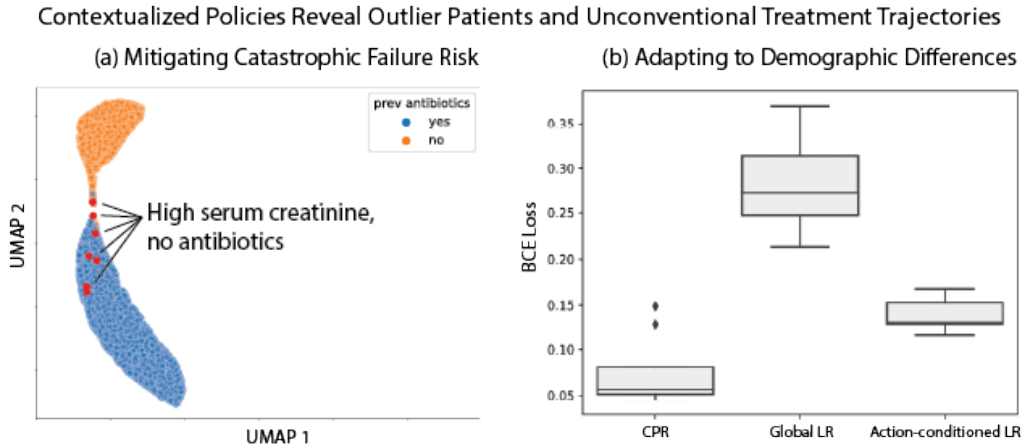
Figure 11 confirms that even with global interpretability,

Contextualized Policies Reveal Outlier Patients and Unconventional Treatment Trajectories

Figure 2. CPR generates decision models for marginal groups with high accuracy. (a) Using only a small subgroup of patients making up 7 observations in the training set, CPR identifies elevated creatinine as a severe risk factor for kidney failure and reassigns patients to a non-antibiotics treatment plan, while these patients would otherwise be likely to receive treatment. (b) For the small subgroup of patients below 20 years of age (with only 9 observations in the held-out set, 44 in train, and 12 in validation), we evaluate global, condition-specific, and contextualized logistic regression (LR) policy models. The global model is a universal LR policy for all patients at all timesteps. In Figure (12 we identify previous antibiotic treatments as a significant driver of policy divergence and train two condition-specific LRs based on whether a patient did or did not receive treatment in the previous timestep. CPR improves significantly over both baselines, capturing highly personalized treatment decisions not accounted for by baseline models. BCELoss is binary cross-entropy loss.

the most important factors in treatment decisions are almost always the most recent observations, confirming the sufficiency and necessity of localized context-specific policies for representing real world behaviors.

Contextualized Policies Reveal Outlier Patients By relating modeling tasks through task-specific contexts, CPR learns to generate context-specific models even when the number of samples per context is as small as one, and similarly generalizes to generate models for unseen contexts. Previously, we assessed these models in aggregate to reveal common treatment trajectories and best practices (Figures 1b, 9). Here, we demonstrate how personalized models also reveal small subgroups of patients or even individuals with outlier policies. In particular, some patient populations have higher dose tolerances and are amenable to more aggressive treatment, while others display rare comorbidities and risk factors prohibit common treatment options. Personalized treatment options are critical, especially when more common treatment plans need to be avoided.

CPR identifies several of these outlier patients when in the MIMIC antibiotic prescription dataset (Fig. 2). First, younger patients often have fewer comorbidities and more robust immune systems, and physicians can be more confident that antibiotics will not impose any adverse side effects if an infection is suspected. We observe that contextualized policies recover this case, and represent treatment for the under 20 age group much more accurately when antibi-

otics are prescribed. Second, elevated creatinine is a rare side-effect of antibiotics but a likely indicator of suboptimal Kidney function and possible Kidney failure (Gounden et al., 2023). CPR identifies that patients with elevated creatinine are immediately removed from antibiotics following an initial prescription, placing them in a treatment cluster characterized by a lack of antibiotics prescription that would otherwise be unlikely for these patients. The patient-specific policies produced by CPR provide a novel view of the treatment process, allowing us to easily identify these rare and outlier effects in terms of context-specific policy parameters and errors, revealing nuances in treatment decisions that were missed by prior works.

4.1.2. ADNI MRI SCANS

Following prior works by (Hüyük et al., 2022) and (Pace et al., 2022), we apply CPR to 1605 patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The canonical task is to predict at each visit whether a Magnetic Resonance Image (MRI) scan is ordered for cognitive disorder diagnosis (noa, 2018). Patient observations consist of the Clinical Dementia Rating (CDR-SB) on a severity scale (normal; questionable impairment; severe dementia) (O'Bryant et al., 2008) and the MRI outcome of the previous visit falling into 4 categories: No MRI scan, below average, average and above average hippocampal volume.

The small number of discrete context features and the checklist-like time-independent diagnostic criteria for
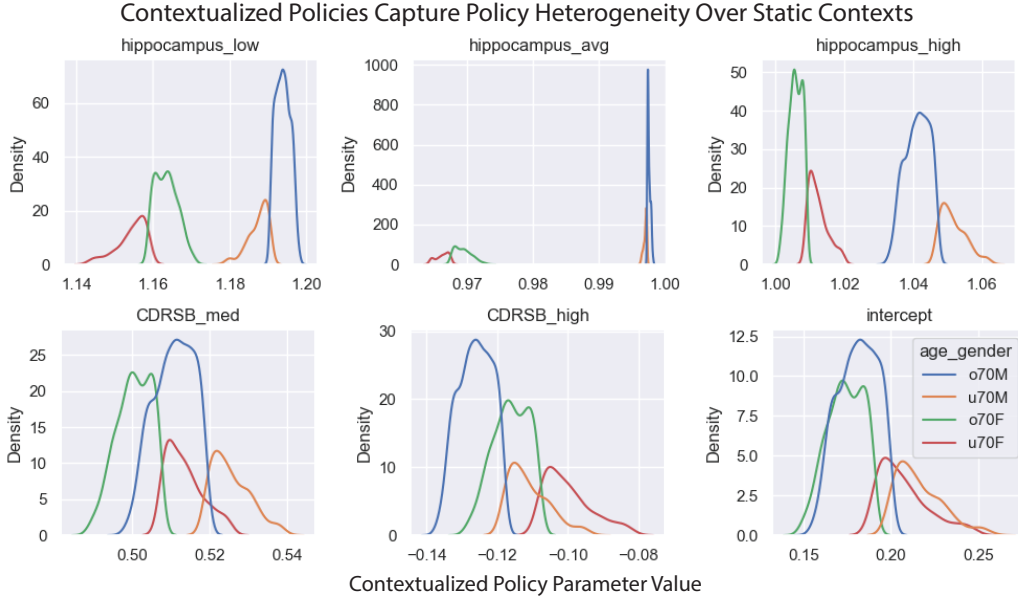
Figure 3. Comparison of the distribution of patient-specific model parameters generated by CPR for predicting MRI prescriptions during Alzheimer's diagnosis. Models are grouped by age and gender in visit $t = 0$ after incorporating age and gender as patient contexts.

Alzheimer's (O'Bryant et al., 2008) provide a limited view of Alzheimer's diagnosis that is unlikely to explain any heterogeneity in clinical decisions. This is reinforced by the fact that a single logistic regression outperforms all interpretable policy baselines (Table 2). We introduce a new condition-specific logistic regression baseline, where we learn a decision model for every set of unique context features at each timestep. Indeed, this condition-specific model performs nearly as well as CPR and black-box models, with an AUROC of 0.71. While CPR and the black-box baselines can still capture dependencies on past actions and observations (Fig. 8, 9), this seems to confer only marginal modeling improvements for the canonical ADNI task.

Instead, we reformulate this canonical task to include a new source of heterogeneity with clinical significance: patient age and gender (Castro-Aldrete et al.). Condition-specific logistic regression and both interpretable baselines (Hüyük et al., 2022; Pace et al., 2022) are unable to model changes in patient-specific policies over continuous static contexts like age, but CPR is able to easily incorporate static as well as dynamic contexts by encoding static contexts into the initial hidden state of the context encoder f.

Figure 3 shows how the estimated policies at $t = 0$ differ between four patient subgroups. We find meaningful heterogeneity in the models generated by CPR, where age dominates CDRSB coefficients and overall intercept, while gender dominates hippocampal volume coefficients. Additionally, static contexts substantially increase the action-matching performance of CPR to 0.763 AUROC.

4.2. Simulations

CPR incorporates both a deep learning component and a statistical modeling component to introduce a novel mechanism of interpretability, the context-specific linear policy. Our approach differs substantially from prior interpretable methods by including a deep learning component. Naturally, we wonder if CPR's explicit linear policy representation is key to its performance and interpretability, or if accurate and robust context-specific linear policies can also be recovered from black-box policy models using post-hoc interpretation methods. To test this, we simulate a heterogeneous, action-dependent Markov decision process (MDP) and evaluate CPR versus black-box baselines on their recovery of simulation parameters: the true action probability and the true coefficients of a context-specific linear policy (Fig. 4). While CPR explicitly generates these context-specific linear coefficients, black-box models implicitly model these coefficients as feature gradients (i.e. linear coefficients in a first-order Taylor expansion). Akin to popular post-hoc interpretability methods like LIME (Ribeiro et al., 2016), we leverage the differentiability of RNNs $\Phi(x_t, h) \rightarrow a_t$ to recover the implicit context-specific linear policies θ.

$$\widehat{\theta} = \frac{\partial}{\partial x_t} \Phi(x_t, h)$$

We generate data from a heterogeneous MDP governed by a known context-specific policy

$$P(a_t = 1 | x_t, x_{t-\tau}, a_{t-\tau}, t) = 1/(1 + \exp(-\theta \cdot x_t + \epsilon_a))$$

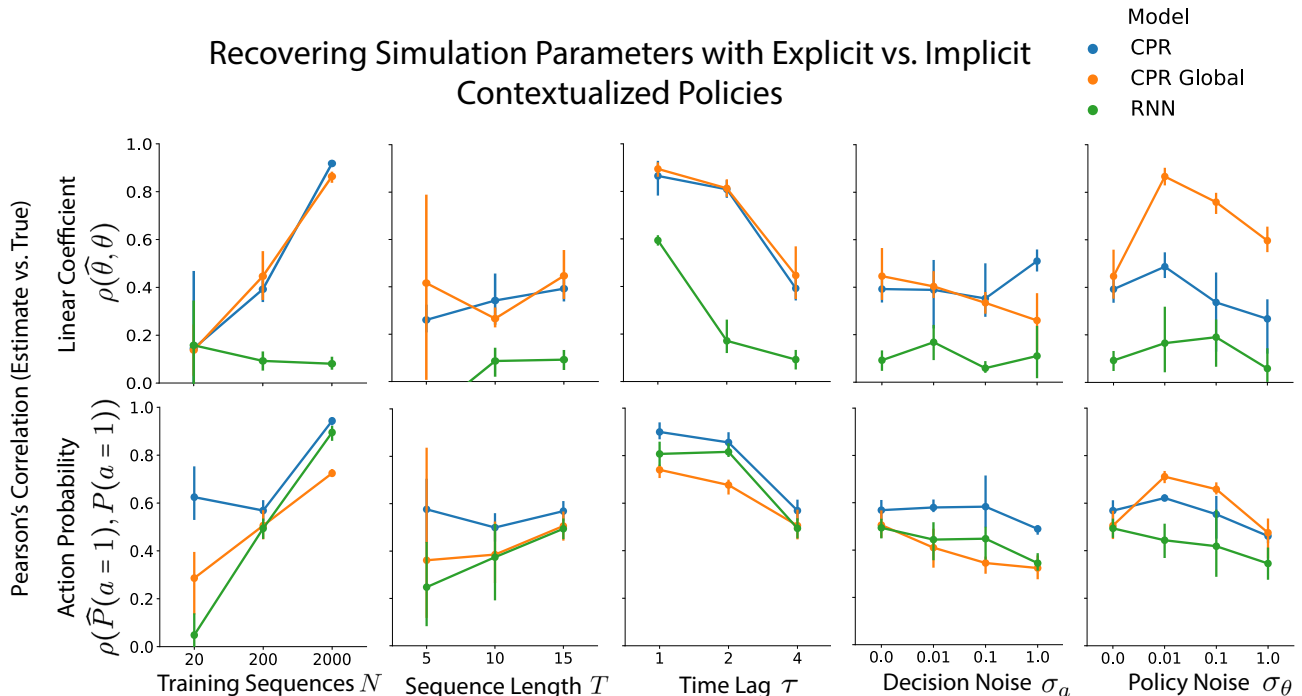$$\theta = x_{t-\tau} \cdot (2a_{t-\tau} - 1) + \frac{t}{T} + \epsilon_\theta$$

Recovering Simulation Parameters with Explicit vs. Implicit Contextualized Policies

Figure 4. Comparing policy models learned by CPR and RNN in terms of the Pearson's correlation between estimated and true action probabilities and context-specific policy coefficients. We choose default simulation arguments $N = 200$, $\sigma_a = 0$, $T = 15$, $\tau = 4$, and $\sigma_\theta = 0$, varying each parameter individually. We hold out 15% of trajectories at random for evaluation. Results are the mean and 95% confidence interval from three randomly initialized and independently simulated data sets.

where T is the sequence length or maximum timestep, and τ is a time lag between the current policy and a dependence on past observations $x_{t-\tau}$ and actions $a_{t-\tau}$. We simulate N total sequences, drawing observations $x_t \sim \mathrm{Unif}[-2, 2]$, policy noise $\epsilon_\theta \sim N(0, \sigma_\theta^2)$, and decision noise $\epsilon_a \sim N(0, \sigma_a^2)$ at each timestep. On a known heterogeneous and action-dependent MDP, CPR's explicit policy representation not only improves its representation of MDP parameters but increases overall performance versus a black-box model with an unstructured policy representation (Fig. 4).

5. Discussion

In this study, we propose contextualized policies as dynamic, interpretable, and personalized linear decision models, each representing a single step in a complex treatment process. By relating individual modeling tasks through patient-specific histories and contexts we avoid the pitfalls of common personalization methods that reduce statistical power (e.g. sample splitting and subpopulation grouping). As a result, CPR matches the performance of black-box models while retaining the interpretability of linear models. Analysis of patient-specifc models generated by CPR reveal rare covariates with outsize effects on treatment decisions, as well as extremely subtle effects in the general population. As a framework, CPR is highly modular and extensible. It

is compatible with any recurrent black-box model or interpretable decision model, and we demonstrate this by developing both locally and globally interpretable context-specific policies under the CPR framework, and applying them with common recurrent network encoders. Together, these policy models demonstrate that local context-specific policies are often sufficient to characterize and interpret real-world behaviors. While we apply CPR in offline and partially observable environments, CPR is directly portable to online policy inference with only subtle training modifications.

Impact Statement

We address fundamental tradeoffs between accuracy and interpretability in imitation learning by developing CPR, a highly accurate and highly interpretable imitation learning method. For biomedical informaticians and statisticians, CPR recovers sample-specific treatment policies, revealing heterogeneity in treatment decisions across patients, predicting personalized policies for individual patients, and identifying key factors differentiating treatment trajectories. For machine learning researchers, CPR is a step toward reinforcement learning agents that explain as they think, and promises a general purpose platform for supporting and improving complex human decisions.

Acknowledgments

References

Dementia: assessment, management and support for people living with dementia and their carers. *National Institute for Health and Care Excellence (NICE) Guideline No. 97*, June 2018. URL https://www.nice.org.uk/guidance/ng97. URL: https://www.nice.org.uk/guidance/ng97.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. Neural additive models: Interpretable machine learning with neural nets. URL https://openreview.net/forum?id=wHkKTW2wrmm.

Ageenko, I. I., Doherty, K. A., and Cleave, A. P. V. Personalized lifetime financial planning tool, June 2010. URL https://patents.google.com/patent/US20100161467/en.

Al-Shedivat, M., Dubey, A., and Xing, E. P. Contextual Explanation Networks, September 2020. URL http://arxiv.org/abs/1705.10301. arXiv:1705.10301 [cs, stat].

Bain, M. and Sammut, C. A Framework for Behavioural Cloning. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]*, pp. 103–129, GBR, January 1999. Oxford University. ISBN 978-0-19-853867-7.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., and Koller, D. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine*, 3(108), November 2011. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.3002564. URL https://www.science.org/doi/10.1126/scitranslmed.3002564.

Bica, I., Jarrett, D., Hüyük, A., and Schaar, M. v. d. Learning "What-if" Explanations for Sequential Decision-Making. March 2021. URL https://openreview.net/forum?id=h0de3QWtGG.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33 (2):155–160, February 2015. ISSN 1546-1696. doi: 10.1038/nbt.3102. URL https://www.nature.com/articles/nbt.3102. Number: 2 Publisher: Nature Publishing Group.

Castro-Aldrete, L., Moser, M. V., Putignano, G., Ferretti, M. T., Schumacher Dimech, A., and Santuccione Chadha, A. Sex and gender considerations in alzheimer's disease: The women's brain project contribution. 15. ISSN 1663-4365. URL https://www.frontiersin.org/articles/10.3389/fnagi.2023.1105620.

Chernozhukov, V. and Hansen, C. Instrumental quantile regression inference for structural and treatment effect models. 132(2):491–525. ISSN 0304-4076. doi: 10.1016/j.jeconom.2005.02.009. URL https://www.sciencedirect.com/science/article/pii/S0304407605000643.

Choi, J. and Kim, K.-E. Inverse Reinforcement Learning in Partially Observable Environments. *Journal of Machine Learning Research*, 12(21):691–730, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/choi11a.html.

Ellington, C. N., Lengerich, B. J., Watkins, T. B., Yang, J., Xiao, H., Kellis, M., and Xing, E. P. Contextualized networks reveal heterogeneous transcriptomic regulation in tumors at sample-specific resolution. URL https://www.biorxiv.org/content/10.1101/2023.12.01.569658v1. Pages: 2023.12.01.569658 Section: New Results.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature21056. URL http://www.nature.com/articles/nature21056.

Fisher, A. J., Medaglia, J. D., and Jeronimus, B. F. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27):E6106–E6115, July 2018. doi: 10.1073/pnas.1711978115. URL https://www.pnas.org/doi/10.1073/pnas.1711978115. Publisher: Proceedings of the National Academy of Sciences.

Frosst, N. and Hinton, G. Distilling a Neural Network Into a Soft Decision Tree, November 2017. URL http://arxiv.org/abs/1711.09784. arXiv:1711.09784 [cs, stat].

Gounden, V., Bhatt, H., and Jialal, I. Renal Function Tests. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. URL http://www.ncbi.nlm.nih.gov/books/NBK507821/.

Guleria, R., Mohan, A., Kulkarni, A., Madan, K., and Hadda, V. Guidelines for Antibiotic Prescription in Intensive Care Unit. *Indian Journal of Critical Care Medicine*, 23(S1):1–63, 2019. ISSN 0972-5229, 1998-359X. doi: 10.5005/jp-journals-10071-23101. URL https://www.ijccm.org/doi/10.5005/jp-journals-10071-23101.

Ha, D., Dai, A., and Le, Q. V. HyperNetworks. URL http://arxiv.org/abs/1609.09106.

Hart, S. Precision Education Initiative: Moving Towards Personalized Education. *Mind, brain and education : the official journal of the International Mind, Brain, and Education Society*, 10(4):209–211, December 2016. ISSN 1751-2271. doi: 10.1111/mbe.12109. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476312/.

Hastie, T. and Tibshirani, R. Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. ISSN 0035-9246. URL https://www.jstor.org/stable/2345993. Publisher: [Royal Statistical Society, Wiley].

Heckman, J., Ichimura, H., Smith, J., and Todd, P. Characterizing selection bias using experimental data, a. URL https://www.nber.org/papers/w6699.

Heckman, J. J., Ichimura, H., and Todd, P. E. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. 64(4):605–654, b. ISSN 0034-6527. doi: 10.2307/2971733. URL https://www.jstor.org/stable/2971733. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Ho, J. and Ermon, S. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Hüyük, A., Jarrett, D., Tekin, C., and Schaar, M. v. d. Explaining by Imitating: Understanding Decisions by Interpretable Policy Learning. February 2022. URL https://openreview.net/forum?id=unI5ucw_Jk.

Jarrett, D., Yoon, J., Bica, I., Qian, Z., Ercole, A., and van der Schaar, M. Clairvoyance: A pipeline toolkit for medical time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xnC8YwKUE3k.

Jeon, W., Seo, S., and Kim, K.-E. A Bayesian Approach to Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635. Number: 1 Publisher: Nature Publishing Group.

Khalili, H., Bairami, S., and Kargar, M. Antibiotics induced acute kidney injury: incidence, risk factors, onset time and outcome. *Acta Medica Iranica*, 51(12):871–878, 2013. ISSN 1735-9694.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *CoRR*, December 2014.

Kostrikov, I., Nachum, O., and Tompson, J. Imitation Learning via Off-Policy Distribution Matching. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hyg-JC4FDr.

Kreif, N., Grieve, R., Radice, R., and Sekhon, J. S. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. 13(2):174–202. ISSN 1572-9400.

Królicka, A. L., Kruczkowska, A., Krajewska, M., and Kusztal, M. A. Hyponatremia in Infectious Diseases-A Literature Review. *International Journal of Environmental Research and Public Health*, 17(15):5320, July 2020. ISSN 1660-4601. doi: 10.3390/ijerph17155320.

Laï, M.-C., Brian, M., and Mamzer, M.-F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *Journal of Translational Medicine*, 18(1):14, December 2020. ISSN 1479-5876. doi: 10.1186/s12967-019-02204-y. URL https://translational-medicine.

biomedcentral.com/articles/10.1186/s12967-019-02204-y.

Lengerich, B., Ellington, C. N., Rubbi, A., Kellis, M., and Xing, E. P. Contextualized machine learning, a. URL http://arxiv.org/abs/2310.11340.

Lengerich, B., Ellington, C., Aragam, B., Xing, E. P., and Kellis, M. NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters, November 2021. URL http://arxiv.org/abs/2111.01104. arXiv:2111.01104 [cs, stat].

Lengerich, B. J., Al-Shedivat, M., Alavi, A., Williams, J., Labbaki, S., and Xing, E. P. Discriminative subtyping of lung cancers from histopathology images via contextual deep learning, b. URL https://www.medrxiv.org/content/10.1101/2020.06.25.20140053v2. ISSN: 2014-0053 Pages: 2020.06.25.20140053.

Lengerich, B. J., Nunnally, M. E., Aphinyanaphongs, Y., Ellington, C., and Caruana, R. Automated interpretable discovery of heterogeneous treatment effectiveness: A COVID-19 case study. pp. 104086, c. ISSN 1532-0464. doi: 10.1016/j.jbi.2022.104086. URL https://www.sciencedirect.com/science/article/pii/S1532046422001022.

Lengerich, B. J., Caruana, R., Nunnally, M. E., and Kellis, M. Death by round numbers: Glass-box machine learning uncovers biases in medical practice. *medRxiv*, pp. 2022–04, 2022.

Lundberg, S. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. ISBN: 1705.07874 Publication Title: arXiv [cs.AI], May 2017. URL http://arxiv.org/abs/1705.07874.

Makino, T. and Takeuchi, J. Apprenticeship learning for model parameters of partially observable environments. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 891–898, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

Masterton, R. G., Galloway, A., French, G., Street, M., Armstrong, J., Brown, E., Cleverley, J., Dilworth, P., Fry, C., Gascoigne, A. D., Knox, A., Nathwani, D., Spencer, R., and Wilcox, M. Guidelines for the management of hospital-acquired pneumonia in the UK: report of the working party on hospital-acquired pneumonia of the British Society for Antimicrobial Chemotherapy. *The Journal of Antimicrobial Chemotherapy*, 62(1):5–34, July 2008. ISSN 1460-2091. doi: 10.1093/jac/dkn162.

McInnes, L., Healy, J., Saul, N., and Großberger, L. UMAP: Uniform manifold approximation and projection. 3(29):861. ISSN 2475-9066. doi: 10.21105/joss.00861. URL https://joss.theoj.org/papers/10.21105/joss.00861.

McKinlay, J. B., Link, C. L., Freund, K. M., Marceau, L. D., O'Donnell, A. B., and Lutfey, K. L. Sources of Variation in Physician Adherence with Clinical Guidelines: Results from a Factorial Experiment. *Journal of General Internal Medicine*, 22(3):289–296, March 2007. ISSN 0884-8734, 1525-1497. doi: 10.1007/s11606-006-0075-2. URL http://link.springer.com/10.1007/s11606-006-0075-2.

Ng, A. Y. and Russell, S. J. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA, June 2000. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-707-1.

Ng, K., Sun, J., Hu, J., and Wang, F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2015:132–136, 2015. ISSN 2153-4063.

O'Bryant, S. E., Waring, S. C., Cullum, C. M., Hall, J., Lacritz, L., Massman, P. J., Lupo, P. J., Reisch, J. S., and Doody, R. Staging Dementia Using Clinical Dementia Rating Scale Sum of Boxes Scores. *Archives of neurology*, 65(8):1091–1095, August 2008. ISSN 0003-9942. doi: 10.1001/archneur.65.8.1091. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409562/.

Ogburn, E. L., Rotnitzky, A., and Robins, J. M. Doubly robust estimation of the local average treatment effect curve. 77(2):373–396. ISSN 1369-7412. doi: 10.1111/rssb.12078. URL https://doi.org/10.1111/rssb.12078.

Pace, A., Chan, A., and Schaar, M. v. d. POETREE: Interpretable Policy Learning with Adaptive Decision Trees. January 2022. URL https://openreview.net/forum?id=AJsI-ymaKn_.

Piot, B., Geist, M., and Pietquin, O. Boosted and reward-regularized classification for apprenticeship learning. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, AAMAS '14, pp. 1249–1256, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-2738-1.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ISBN: 1602.04938 Publication Title: arXiv [cs.LG], February 2016. URL http://arxiv.org/abs/1602.04938.

Royal Society (Great Britain) and Royal Society (Great Britain) Staff. *Machine Learning: The Power and Promise of Computers That Learn by Example*. Royal Society, 2017. ISBN 978-1-78252-259-1.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences, a. URL http://arxiv.org/abs/1704.02685.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences, b. URL http://arxiv.org/abs/1605.01713.

Sun, W., Venkatraman, A., Gordon, G. J., Boots, B., and Bagnell, J. A. Deeply AggreVaTeD: Differentiable Imitation Learning for Sequential Prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3309–3318. PMLR, July 2017. URL https://proceedings.mlr.press/v70/sun17d.html. ISSN: 2640-3498.

Verma, A., Murali, V., Singh, R., Kohli, P., and Chaudhuri, S. Programmatically Interpretable Reinforcement Learning. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5052–5061. PMLR, 2018. URL http://proceedings.mlr.press/v80/verma18a.html.

Westert, G. P., Groenewoud, S., Wennberg, J. E., Gerard, C., DaSilva, P., Atsma, F., and Goodman, D. C. Medical practice variation: public reporting a first necessary step to spark change. *International Journal for Quality in Health Care*, 30(9):731–735, November 2018. ISSN 1353-4505, 1464-3677. doi: 10.1093/intqhc/mzy092. URL https://academic.oup.com/intqhc/article/30/9/731/4990397.

Yau, H., Russell, C., and Hadfield, S. What Did You Think Would Happen? Explaining Agent Behaviour through Intended Outcomes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18375–18386. Curran Associates, Inc., 2020.

A. Appendix

Code for data processing, model training, and figure generation is available on GitHub [1]. All model embeddings are done with UMAP (McInnes et al.).

A.1. Data

Each dataset is split up into a training set (70% of patients), validation set (15% of patients) for hyperparameter tuning, and test set (15% of patients) to report model performance.

A.1.1. ADNI

We follow the task described by (Pace et al., 2022) and (Hüyük et al., 2022), taking the preprocessing code from (Hüyük et al., 2022). Overall, the dataset contains patients with at least two and a median of three visits. We exclude patients without CDRSB measurements in one of their visits.

A.1.2. MIMIC

We follow the task of predicting antibiotic prescription in the intensive care unit over up to 6 timesteps as described by (Pace et al., 2022) adapting the preprocessing code provided by (Jarrett et al., 2021). Patient trajectories with missing values in one of the observed features or non-consecutive measurements, patients above 100 years or below 1 year of age, and short stays below 3 days were eliminated. We end up with 4195 patients in our dataset. Input features were standardized and the hemoglobin feature used in previous work was removed since it is highly correlated with the hematocrit feature (> 0.95). See Table 3 for the results using all 8 input features. Since measurements are taken as averages over each day in the ICU, we use the measurements of the previous day as observations x_t to predict a_t ensuring that we don't predict actions from measurements taken after the patient was treated with antibiotics earlier in the day.

A.2. Implementation

CPR implements g as a recurrent neural network (either vanilla RNN or LSTM) with a hidden state of dimension k and h as a neural network with one hidden layer, again of size k, mapping to the output parameters θ. The initial context c_0 at timestep $t = 0$ is set to all zeros.

Black-Box RNN and LSTM are implemented similarly, with k being the dimensionality of the hidden state and one hidden layer of size k directly mapping to the predicted actions a_t. The black-box model predicts $P(a_t|x_t, h_t)$ by taking $[x_t, a_{t-1}]$ as input at each timestep and is optimized using the binary cross entropy loss.

[1] https://github.com/JADEUSC/contextualized_policy_recovery

| Class | Algorithm | MIMIC Antibiotics | | |
		AUROC	AUPRC	Brier ↓
Interpretable	Logistic regression	0.60	0.79	0.20
	INTERPOLE †	NR	NR	NR
	INTERPOLE ‡	0.65	NR	0.21
	POETREE ‡	0.68	NR	0.19
	CPR-RNN (ours)	0.79	0.88	0.15
	CPR-LSTM (ours)	0.80	0.88	0.15
Black-Box	RNN	0.81	0.89	0.14
	LSTM	0.82	0.89	0.14

Table 3. Action-matching performance of different policy learning algorithms on the MIMIC antibiotics task using all 8 input features. Listed performance of INTERPOLE and POETREE are from prior reports of state-of-the-art performance on these canonical datasets: † reported in (Hüyük et al., 2022), ‡ reported in (Pace et al., 2022). NR: No values reported.

| Class | Algorithm | ADNI MRI scans | | MIMIC Antibiotics | |
		λ	k	λ	k
Interpretable	CPR-RNN (ours)	0.0001	32	0.0001	32
	CPR-LSTM (ours)	0.001	64	0.0001	32
Black-Box	RNN	-	64	-	32
	LSTM	-	64	-	64

Table 4. Hyperparameters chosen for different models.

We also introduce CPR-global as a linear combination of all past predictions. While this provides global interpretability, it can also accumulate errors over time. This error compounding can limit performance compared to CPR. We introduce the hyperparameter α to partially address this and allow a user-defined rate of "forgetting" in the model.

A.3. Training

We train all models using the Adam optimizer (Kingma & Ba, 2014) and early stopping on the validation set. The initial learning rate chosen for CPR is 5e-4 and 1e-4 for the baseline RNNs. We select the dimensions of the hidden state for both CPR and the baseline RNNs from $[16, 32, 64]$. For CPR, λ is chosen from $[0.0001, 0.001, 0.01, 0.1]$. The batch size is selected as 64 for all models. Table 4 shows the optimal hyperparameters chosen based on the validation set performance.

A.4. Additional Experiments

A.4.1. SIMULATIONS

Figure 6 shows that CPR is able to recover context-dependent threshold decision boundaries where an agent takes an action if the observed value is above or below a certain threshold. Here x_t is sampled as $x_t \sim \text{Unif}[0, 1]$.

We further simulate dynamic, history-dependent decision policies in a homogeneous MDP and evaluate CPR based

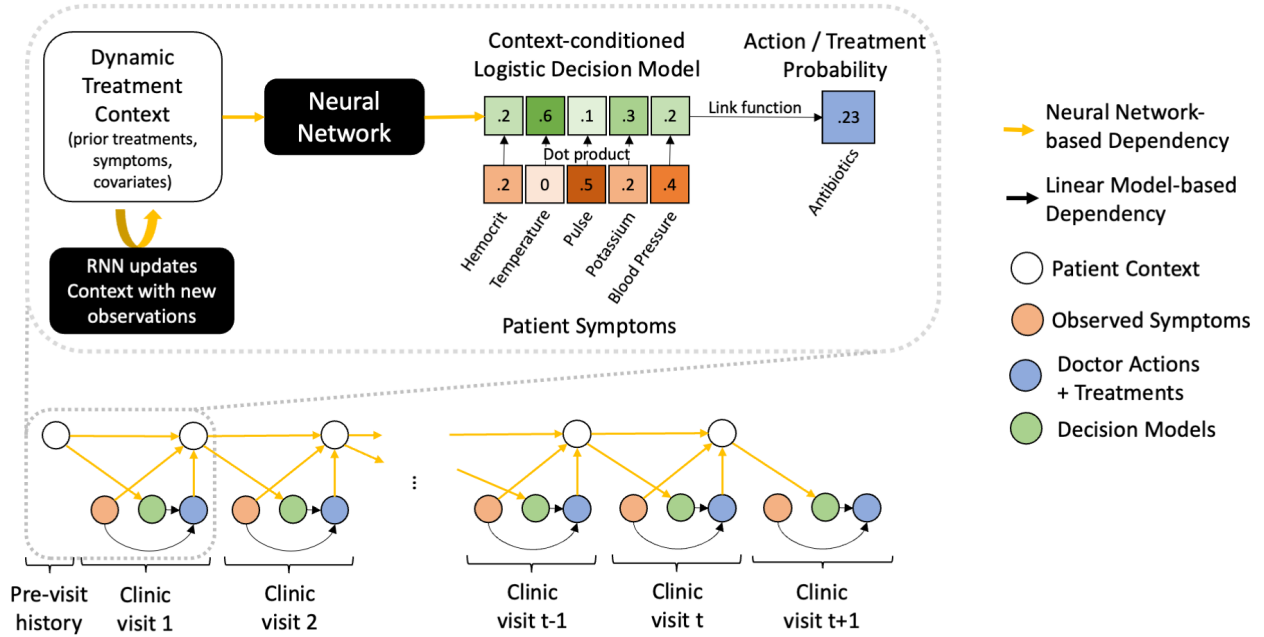## Modeling Medical Decisions with Dynamic Treatment Context

Figure 5. CPR uses dynamic treatment contexts to generate the agent's decision model at each timestep. Decision models are a context-specific weighted combination of observed features. With these context-specific linear decision models, CPR achieves exact model-based interpretability without sacrificing representational capacity.
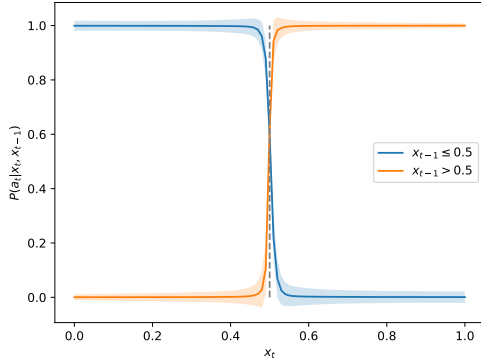

Figure 6. CPR recovers hard decision boundaries, generated by rule-based decision making over time. Here, when the previous observation $x_{t-1} < 0.5$, the current action is taken if $x_t < 0.5$. If $x_{t-1} \geq 0.5$, the current action is reversed. The probabilistic models of these boundaries align closely with the true step function.

it's recovery of true simulated true parameters in these policies. We generate $n = 2000$ patient trajectories of length $T = 9$ time steps. At each time step, we sample a random observation variable $x_t \sim \text{Unif}[-1, 1]$ and an agent which takes an action $a_t \in [0, 1]$. This action is determined by a true observation-to-action mapping function which depends on the observed history and total treatment time.

$$P(a_t | x_t, x_{t-1}) = \sigma(w^t(x_{t-1}) * x_t + b^t(t)), \qquad (1)$$

where $w^t(x_{t-1}) = 4x_{t-1}$ and $b^t(t) = \frac{t-5}{4}$. We simulate a true contextual policy and assess recovery of simulation parameters with CPR and a RNN. The probability of taking an action at each time t depends on both the absolute point in the time series, as well as the history of observations immediately preceding any time point. We design this simulation as a succinct way to demonstrate how simple decision models which depend on both absolute and relative effects over a time series often combine to create a universal policy that is exceptionally complex and difficult to capture in a single model that is invariant to time or context. We evaluate CPR based on it's ability to imitate the observed time-dependent and context-dependent actions and recover the simulated true parameters.

Our method is able to recover the true model parameters (marked in red) as shown in Figure 7 with a slight bias in the upper and lower value range.

A.4.2. ADNI MRI SCANS

Bootstrapped Results We run 10 bootstrap runs to get confidence estimates for model performance on the ADNI (Table 5) dataset. Each bootstrap sample is randomly split
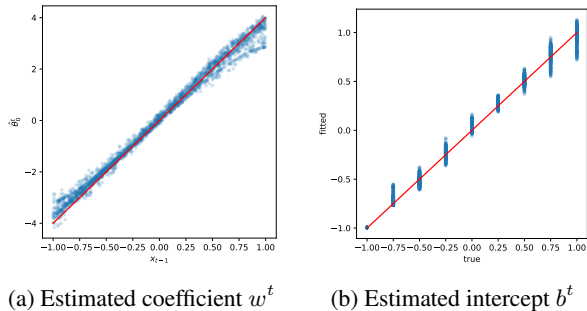
(a) Estimated coefficient w^t

(b) Estimated intercept b^t

Figure 7. CPR recovers true policy coefficients in a homogeneous MDP

into a training, validation and test set.

Class	Algorithm	ADNI MRI scans		
		AUROC	AUPRC	Brier ↓
Interpretable	Logistic regression	0.66 ± 0.03	0.86 ± 0.01	0.16 ± 0.01
	INTERPOLE †	0.60 ± 0.04	0.81 ± 0.08	0.17 ± 0.05
	INTERPOLE ‡	0.44 ± 0.04	0.75 ± 0.09	0.19 ± 0.07
	POETREE ‡	0.62 ± 0.01	0.82 ± 0.01	0.18 ± 0.01
	CPR-RNN (ours)	$\mathbf{0.72 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.15 \pm 0.01}$
	CPR-LSTM (ours)	$\mathbf{0.72 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.15 \pm 0.01}$
Black-Box	RNN	0.72 ± 0.02	0.88 ± 0.02	0.15 ± 0.01
	LSTM	0.71 ± 0.02	0.88 ± 0.02	0.15 ± 0.01

Table 5. Action-matching performance of different policy learning algorithms on the ADNI MRI scans task. Bolded values in each column denote the best performance of any interpretable model. Listed performance of INTERPOLE and POETREE are from prior reports of state-of-the-art performance on these datasets: † reported by (Hüyük et al., 2022), ‡ reported by (Pace et al., 2022).

Figure 8 shows how different contexts influence the agents decision function over time. Most notably, a high CDRSB value during the first two visits decreases the probability of ordering an MRI since this can be already seen as a strong indicator of dementia, making a scan less informative (noa, 2018). Patients A and B share the same decision model in $t = 1$ since they both show medium CDRSB and avg hippocampal volume in $t = 0$. Afterward, their decision function differs in $t = 2$. Patient A's hippocampal volume was measured as "low" in $t = 1$ leading to a lower overall probability of ordering an MRI in $t = 2$ indicated by a lower intercept and CDRSB coefficients. In contrast, Patient B was diagnosed with an "avg" hippocampal volume in visit $t = 1$. A low hippocampal volume can again be seen as a strong indicator of dementia making a scan less informative. Patient C, in contrast, is diagnosed with medium CDRSB and high hippocampal volume throughout all visits. This increases the probability of MRI (higher intercept) since there is no clear indication that would make an MRI obsolete.

Static Contexts unrolled trough time Static Contexts such as age and gender do not only influence the observation-to-

action mapping at timestep $t = 0$ but uncover heterogeneity across subgroups throughout time. Figure 9 shows that the difference in the intercept coefficient between patients below 70 and above 70 years of age widens over time for both patients A and B indicating that ordering an MRI is less likely for older patients. We can also see a slight difference in the CDRSB_med coefficient with it being slightly negative for old patients, reducing the probability of getting ordering an MRI if an old patient in this patient group scored a medium result, compared to young patients.

A.4.3. MIMIC ANTIBIOTICS

Bootstrapped Results We run 10 bootstrap runs to get confidence estimates for model performance on the MIMIC (Table 6) dataset. Each bootstrap sample is randomly split into a training, validation and test set.

Class	Algorithm	MIMIC Antibiotics		
		AUROC	AUPRC	Brier ↓
Interpretable	Logistic regression	0.57 ± 0.03	0.80 ± 0.03	0.20 ± 0.01
	INTERPOLE †	NR	NR	NR
	INTERPOLE ‡	0.65	NR	0.21
	POETREE ‡	0.68	NR	0.19
	CPR-RNN (ours)	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	0.14 ± 0.01
	CPR-LSTM (ours)	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.14 \pm 0.00}$
Black-Box	RNN	0.83 ± 0.01	0.90 ± 0.01	0.13 ± 0.01
	LSTM	0.84 ± 0.01	0.91 ± 0.01	0.13 ± 0.00

Table 6. Action-matching performance of different policy learning algorithms on the MIMIC antibiotics task. Bolded values in each column denote the best performance of any interpretable model. Listed performance of INTERPOLE and POETREE are from prior reports of state-of-the-art performance on these canonical datasets: † reported by (Hüyük et al., 2022), ‡ reported by (Pace et al., 2022). NR: No values reported.

Model Coefficients Figure 12 shows the heterogeneity in estimated model coefficients. The main drivers of heterogeneity are the intercept, potassium and creatinine.

Treatment Trajectories To uncover typical treatment trajectories, we look at clusters of coefficients over 4 timesteps. We cluster patients into 5 subgroups using hierarchical clustering based on silhouette score (0.566) as seen in Figure 10. Figure 10 shows the two largest clusters containing patients that get antibiotics throughout their stay (cluster 1, 1707 patients) and patients that never get antibiotics (cluster 5, 384 patients). The remaining three groups are plotted in Figure 13. Patients that get treatment for the first two days fall into cluster 3 (232 patients). We can see that their treatment parametrization changes significantly after treatment is stopped. Cluster 4 contains patients that get treated for the first day only (240 patients). Both patients in cluster 3 and 4 share similar decision models in $t = 1$ since both were treated in $t = 0$, it diverges in $t = 2$ after treatment is stopped for one group and share the decision parametrization in $t = 3$ after both groups were not treated in $t = 2$. The remaining
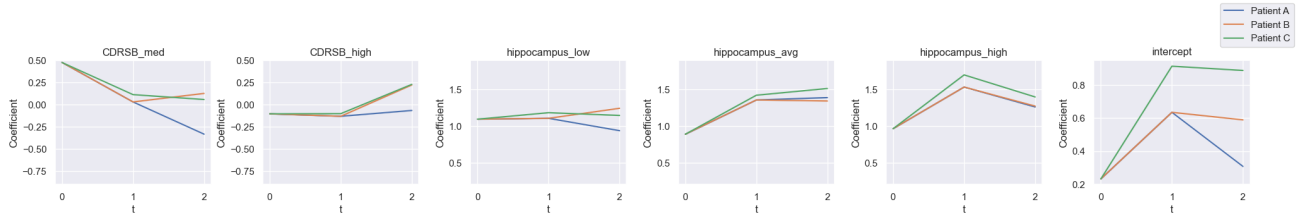
Figure 8. Estimated decision model coefficients over 3 timesteps for 3 patients with representative ADNI contexts. No confidence intervals are available, as the discrete diagnostic features in ADNI produce only a discrete number of possible trajectories.
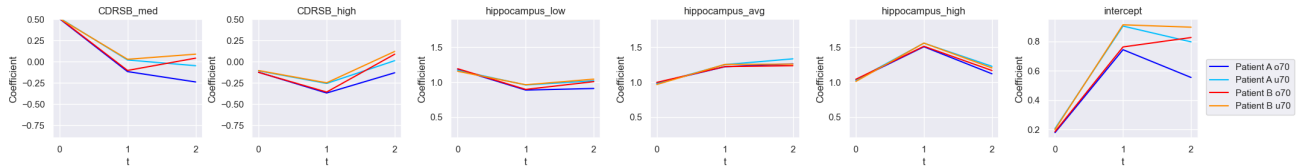

Figure 9. Average estimated coefficients for different ADNI contexts, patients over 70 vs patients under 70. Patient A and B from Figure 8 above. Older patients are less likely to receive MRIs for both patient groups. Static context is age and gender.

patients fall into cluster 4 (375 patients).

Global interpretability of CPR The coefficients β of historical features and actions in previous time steps 0-3 and θ of current features at current time step 4 for predicting current treatment at time step 4 are shown in Fig. 11.

Figure 10. MIMIC patient trajectory clusters, produced by concatenating decision models over 4 consecutive time points into a trajectory matrix and embedding them with UMAP.
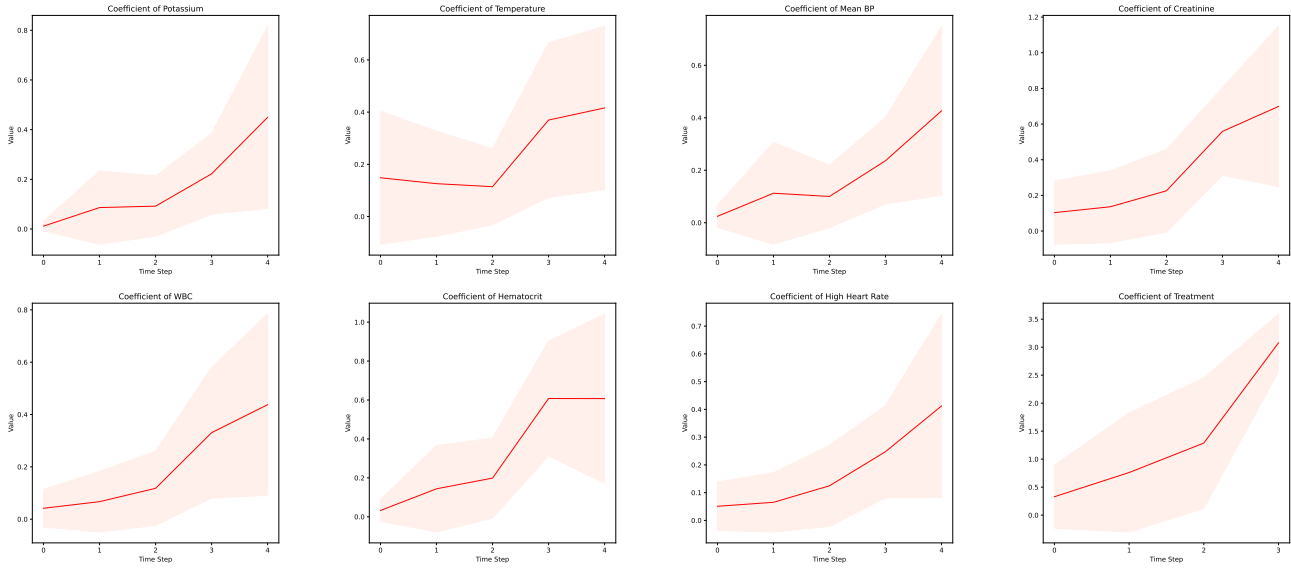
Figure 11. The magnitude of coefficients (linear effects) on each feature within CPR global policies at timestep 4 on MIMIC data. In CPR global, each of CPR's predictions can be decomposed as a linear combination of all observed features and actions, rather than only interpreting current features.
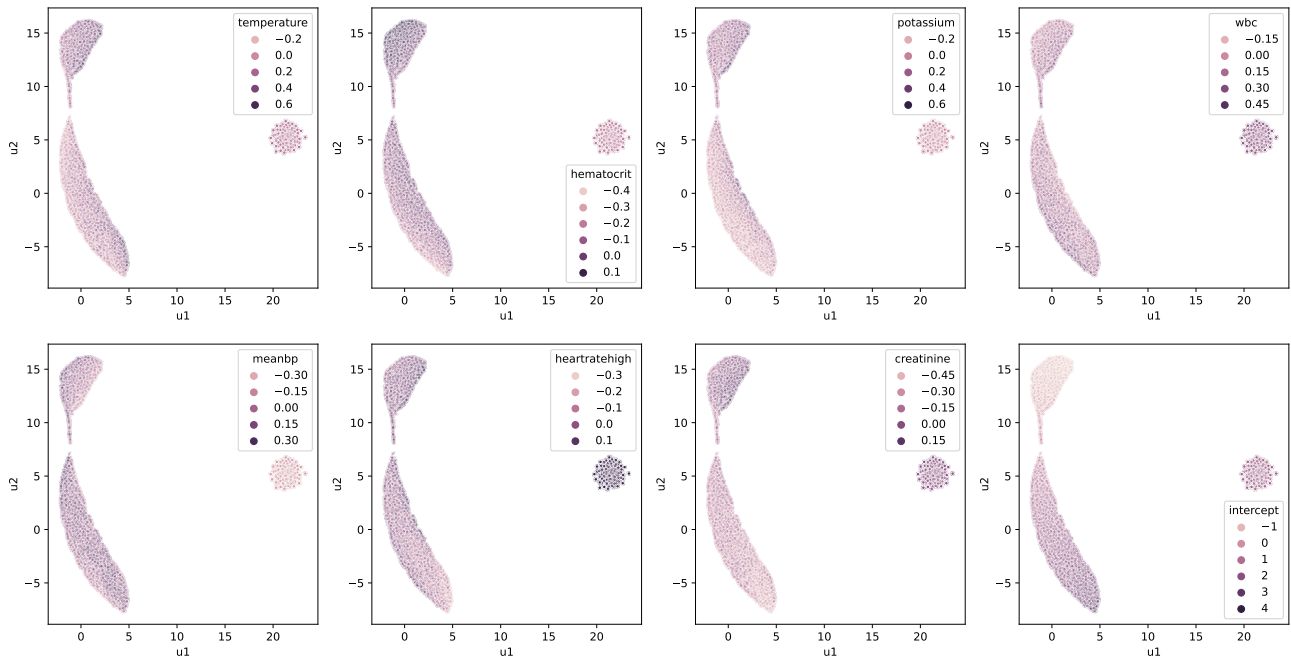

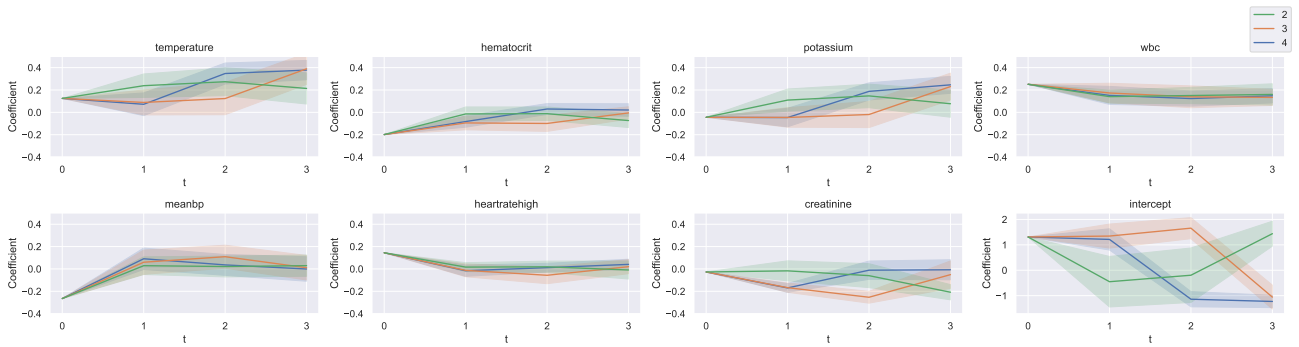*Figure 12.* Model coefficient embeddings for MIMIC policy models.

Figure 13. Estimated model parameters for three remaining model trajectories (first two in Fig. 1c). Error bars are coefficient standard deviations across patients in each trajectory group and time point.
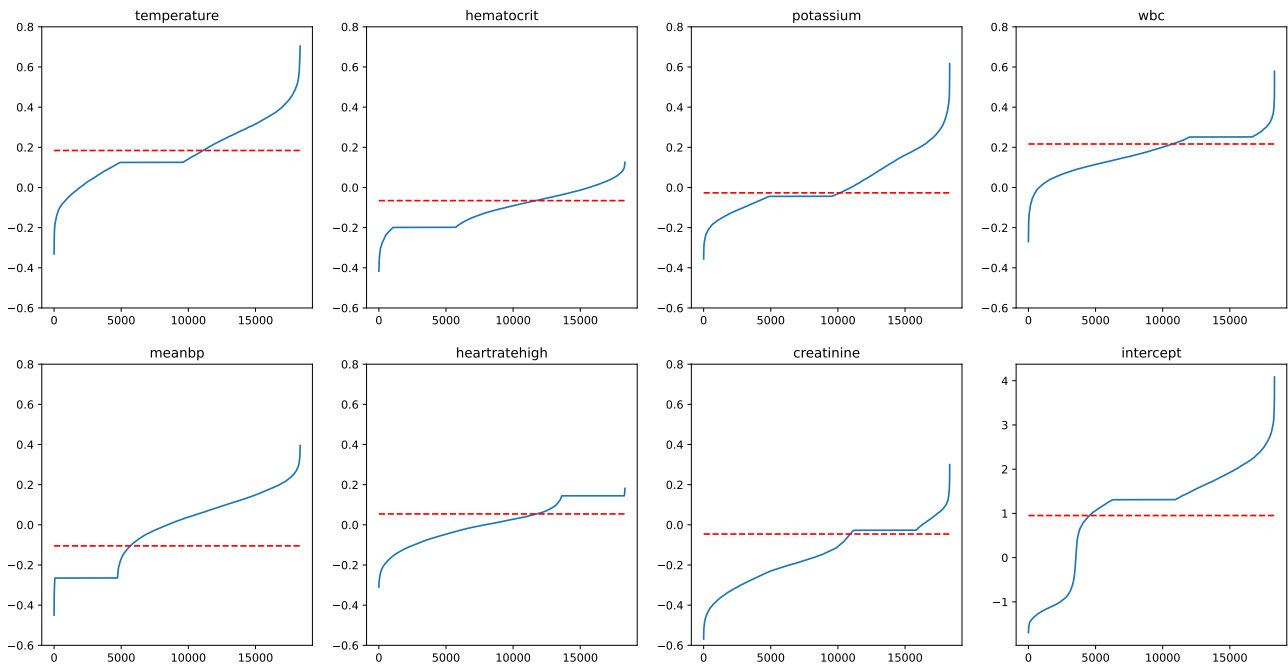

Figure 14. Coefficient values of antibiotic prescription models (MIMIC) parametrized by context vs parameters of global model (dashed red line)