Counterfactually Guided Policy Transfer in Clinical Settings

Taylor W. Killian,
University of Toronto, Vector Institute
twkillian@cs.toronto.edu

Marzyeh Ghassemi
University of Toronto, Vector Institute
marzyeh@cs.toronto.edu

Shalmali Joshi
Vector Institute
shalmali@vectorinstitute.ai

Abstract

Reliably transferring treatment policies learned in one clinical environment to another is currently limited by challenges related to domain shift. In this paper we address off-policy learning for sequential decision making under domain shift—a scenario susceptible to catastrophic overconfidence— which is highly relevant to a high-stakes clinical settings where the target domain may also be data-scarce. We propose a two-fold counterfactual regularization procedure to improve off-policy learning, addressing domain shift and data scarcity. First, we utilize an informative prior derived from a data-rich source environment to indirectly improve drawing counterfactual example observations. Then, these samples are then used to learn a policy for the target domain, regularized by the source policy through KL-divergence. In simulated sepsis treatment, our counterfactual policy transfer procedure significantly improves the performance of a learned treatment policy.

1 Introduction

With continued progress in the development of machine learning algorithms there is increasing interest in deploying models in complex, safety-critical clinical settings [18], including reinforcement learning (RL) applied to sequential decision making [20, 31, 50, 67]. However, domain shift between training and deployment environments [44, 66] presents a significant challenge largely unaddressed in recent RL work within clinical settings. Consider a setting where a treatment policy is developed at a large research hospital (a *source* environment) over several years to aid in acute patient care [19]. Now, consider a small clinical facility (a *target* environment) that caters to a subpopulation within the same city. Within the context of *off-policy* RL [9, 61], it is natural to consider whether the policy learned in the source environment can be reliably deployed in a separate target environment.

Possible roadblocks to effective *transfer* of treatment policies is an inconsistency in measurements and different care practices between environments for patients with comparable health conditions [59, 60] requiring some adaptation of the transferred source policy. We specifically consider when effects of any decision may be non-deterministic or partially observed [27]. Augmenting this underlying process with causal assumptions [46, 47] has recently shown promise in characterizing the uncertainty and unreliability of non–deterministic discrete states within the same environment [7]. Such augmentation has also led to promising improvements in off-policy search by sampling trajectories under a counterfactual distribution.

Several challenges may arise when few samples are observed in the target environment with unintended consequences. Naively learned policies in such cases can significantly overfit and fail to generalize [16], with potentially drastic consequences in the presence of hidden confounding [29]. Also, a policy learned from limited data will produce estimation errors, providing unreliable policies

Preprint. Under review.

that may lead to mistimed or inappropriate interventions with adverse consequences [4, 40, 65]. This suggests that policy learning can potentially benefit from transferring policies learned in data-rich source environments. To do this reliably, we incorporate causal assumptions into the POMDP framework when transferring a trained policy to guide treatment strategies in a target environment.

We propose a two-fold counterfactual regularization procedure for off-policy learning using elements from a source environment that can be feasibly shared in clinical settings. First, an informative prior over the observed transitions from a data-rich source environment is used to improve counterfactual rollouts in a data-scarce environment modeled as a Gumbel-Max SCM [43]. Within this sampling procedure, we augment topdown gumbel-sampling [37] with a new mixture prior that accounts for uncertainty over transition estimates in the target environment, while maintaining counterfactual stability (Sec. 2.1). Next, we combine these counterfactual trajectories with explicit regularization of the learned *target* policy by the optimal *source* policy via KL-divergence aggregation [24]. We introduce these two components of counterfactual regularization in Section 3 and outline how they combine to form a counterfactually guided policy learning algorithm in a data-scarce target environment. This overall procedure, driven by counterfactual regularization, is used to demonstrate significant performance improvement and reduced overconfidence in off-policy evaluation measures in Section 4. These performance gains are shown across (distributionally) disparate clinical environments in a simulated sepsis treatment scenario where the treated populations differ in terms of their unobserved diabetic phenotype composition.

2 Background and Related Work

In this paper we propose an approach for off-policy RL within a counterfactual framework for transfer learning. In a data-scarce environment, transition estimates are prone to error [14, 39] limiting the effectiveness of counterfactual inference and further compounding challenges of off-policy learning. Transfer learning within RL can be used accelerate or improve policy learning in an unknown target environment [33, 62]. Individual observations or trajectories from a source environment have been previously used to directly infer rewards [6, 32] or otherwise accelerate policy convergence [64]. However, the difficulty of the clinical setting that motivates our work is partly characterized by hidden confounding such that observed transition statistics between environments may differ although the dynamics do not. To address this, we propose a novel way to incorporate inductive bias using the source environment's transition statistics indirectly—through counterfactual inference—to leverage sub-spaces of observations that may not be in the target environment.

Causal inference may be used to formalize counterfactual investigations of the change to underlying data distributions following prospective or retrospective interventions. These tools have demonstrated benefits when addressing domain shift in supervised learning problems [2, 51] as well as for policy reuse across multiple environments in simple bandit [5, 34, 35] and in multi-agent settings [15]. Some of these approaches rely on notions of invariance between learned features or actions and the correct outcome. These invariances guard against miscalibrated model performance under domain shift or allow the extension of trained models to new environments. In the context of sequential decision making, these concepts can be thought of as safeguards against potentially harmful or unreasonable action selection. However, some these methods may rely on online data collection which is unsuitable for the off-policy clinical setting that motivates this paper. In recent years, these concepts have been found to be useful in measuring the quality of policies learned from observational data [3, 49] (including within partially observable environments [63]) as well as establishing generalization bounds when estimating causal effects in decision making problems [26].

Healthcare is an intrinsically sequential process where a clinician proposes a treatment, observes the patient response and then adapts or continues their treatment decisions accordingly. In this way the RL paradigm appears particularly promising toward the development of optimal treatment strategies [67], despite challenges presented by retrospective data that is likely confounded in several unobservable ways [20]. By applying structured transfer learning in healthcare settings, there is potential to develop personalized treatment strategies [30, 41, 53] and also work to mitigate health inequities [57] and a lack of model generalization across clinical environments [44, 66]. There has been extensive work applying RL to healthcare problems such as schizophrenia [58], HIV [11] as well as acute care settings such as sepsis [31, 50] and mechanical ventilation [48] to determine optimal treatment strategies. There has also been extensive efforts made to allow for coherent off-policy evaluation of such policies since they cannot be directly tested [17, 21, 28]. Dann et al. [10] developed an evaluation method that introspectively reports policy performance to provide an indication of how reliable a learned treatment strategy may be based on the expected outcome. Counterfactual reasoning in this context has been used to infer improved individualized treatment policies in healthcare by accounting for hidden confounding through missing data [45] or long-term effects of action selection [54] when developing personal treatment policies. Yet each of these approaches rely on sufficiently large and diverse training data to provide adequate counterfactuals. Our proposed transfer framework seeks to address learning in data-limited environments by leveraging learned quantities from a data-rich source environment.

Our work specifically relies on inducing bias [25] indirectly by leveraging causal frameworks to incorporate an informative prior from the source environment in a partially observed sequential decision making setup. In particular, we focus on improving transition dynamics in a data-scarce environment by leveraging counterfactual rollouts perturbed by source exogenous noise variables. Additionally, estimating value functions to guide policy development is prone to overestimation [22] and high variance [52]. Various efforts have sought to regularize the policy learning process to maintain stability as policies improve. Recent trust region approaches to policy optimization [55, 56, 1] use a KL-divergence constraint to keep the updated policy "close" to the previous best policy. We use a similar KL-divergence mechanism albeit to directly constrain the target policy to maintain features of the source policy during learning via a form of regularized policy iteration [12, 13]. To the best of our knowledge, our work is the first to leverage a form regularized policy iteration for transfer learning.

2.1 Causal modeling in Partially Observed Environments

We model our sequential decision making system as a POMDP formalized by a Structural Causal Model (SCM) [7, 46]. A structural causal model $\mathcal{M}$ describes the causal mechanisms driving a system with an ordered triple $\langle \mathbf{U}, \mathbf{X}, \mathbf{F} \rangle$: namely, the set of independent exogenous random variables $\mathbf{U}$, the endogenous variables $\mathbf{X}$ and, the set of functions $\mathbf{F}$ that govern the causal mechanisms. $\mathbf{PA}_i$ are the parents of X_i in the associated causal DAG $\mathcal{G}$. Additional background on SCMs is provided in Appendix A.

We consider a finite-state, finite-action episodic POMDP with discounted rewards. States as $S_t \in \mathcal{S}$, actions as $A_t \in \mathcal{A}$, for $t = \{0, 1, \ldots, T\}$. Let the reward at time t for be denoted by $R_t = \mathcal{R}(S_t, A_t)$, and observations by $O_t \in \mathcal{O}$, which include the rewards. The observed history up to time t ($H_t = \{O_1, A_1 \ldots, A_{t-1}, O_t\}$) can be used to infer the next action. Let $\mu(A_t|H_t)$ represent the stochastic policy with the corresponding observed trajectories $\tau = (S_1, A_1, O_1, \ldots, S_{t-1}, A_{t-1}, O_t)$ with density $p^{\mu}(\tau)$. A POMDP can be represented as an SCM by expressing corresponding conditional distributions, e.g. state-transitions $P(S_{t+1}|S_t, A_t)$, as deterministic functions $s_{t+1} \coloneqq f(s_t, a_t, U_{s,t})$ with exogenous noise $U_{s,t}$ [7, 46]. We denote the intervention of executing a policy π instead of μ by $I(\mu \to \pi)$ and the corresponding SCM as $\mathcal{M}^{do(I(\mu \to \pi))}$. We denote the source environment with superscript `S` and the target environment as `T`. Additionally, in the empirical demonstration of our contributions, the state space $\mathcal{S}$ differs from $\mathcal{O}$ only in the unobserved latent dynamics. As such, throughout this paper we use observation and state interchangeably, denoted by S_t (and s_t).

The Gumbel-Max SCM: Oberst and Sontag [43] introduced the Gumbel-Max SCM, in which all nodes $\mathbf{X}$ are discrete random variables. Given independent Gumbel variables $\mathbf{g} = \{g_1, g_2, \ldots, g_k\}$, the causal mechanisms are given by: $X_i \coloneqq \arg\max_j \log p(X_i = j|\mathbf{PA}_i) + g_j$, effectively embedding the Gumbel-Max Trick [23, 36–38]. The properties of this model ensure that counterfactual queries will always preserve the observed outcomes (defined as *counterfactual stability*). The Gumbel-Max SCM was also used to develop a counterfactual off-policy evaluation procedure (CF-PE) that enables introspection into policies within partially observed off-policy settings. CF-PE is used throughout this paper to compare and analyze all learned policies. This allows for analysis based on specific decisions rather than solely on outcomes [10]. However practical challenges due to sample size or other factors of variation in their learning environment were not investigated. Building from the Gumbel-Max SCM, we incorporate inference over a source environment for counterfactual trajectory generation to aid policy learning in a data-sparse target environment.

3 Counterfactually Guided Policy Transfer

In this section we present Counterfactually Guided Policy Transfer (CFPT) which involves two-fold counterfactual regularization, leveraging a data-rich source environment to improve policy learning in a data-scarce target environment that is partially observed with hidden confounding. Patients in both environments are presumed to have comparable health conditions but the proportions of patient

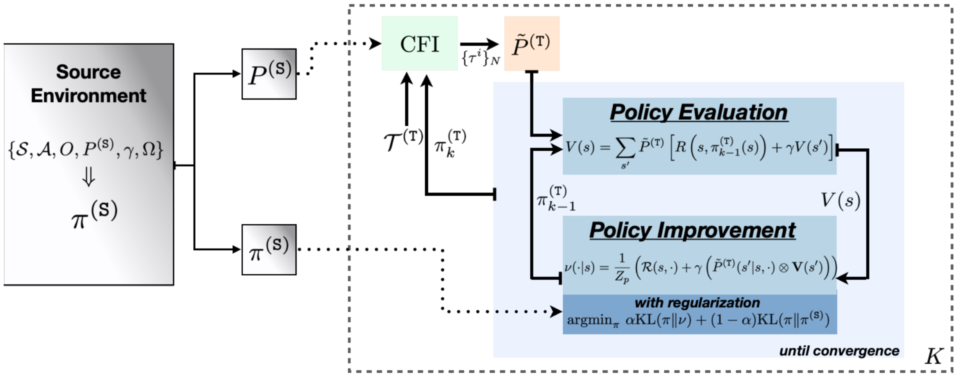
Source Environment
$\{\mathcal{S}, \mathcal{A}, O, P^{(\mathtt{S})}, \gamma, \Omega\}$
$\Downarrow$
$\pi^{(\mathtt{S})}$
$P^{(\mathtt{S})}$
$\pi^{(\mathtt{S})}$
CFI
$\{\tau^i\}_N$
$\tilde{P}^{(\mathtt{T})}$
$\mathcal{T}^{(\mathtt{T})}$
$\pi_k^{(\mathtt{T})}$
Policy Evaluation
$V(s) = \sum_{s'} \tilde{P}^{(\mathtt{T})} \left[R\left(s, \pi_{k-1}^{(\mathtt{T})}(s)\right) + \gamma V(s') \right]$
$\pi_{k-1}^{(\mathtt{T})}$
$V(s)$
Policy Improvement
$\nu(\cdot|s) = \frac{1}{Z_p}\left(\mathcal{R}(s,\cdot) + \gamma\left(\tilde{P}^{(\mathtt{T})}(s'|s,\cdot) \otimes \mathbf{V}(s')\right)\right)$
with regularization
$\text{argmin}_\pi\ \alpha \text{KL}(\pi \| \nu) + (1-\alpha)\text{KL}(\pi \| \pi^{(\mathtt{S})})$
until convergence
K

Figure 1: Graphical overview of counterfactually guided policy transfer as discussed in Sections 3.4 and outlined in Algorithm 4. Elements from the source environment are used to improve counterfactual inference (CFI) and regularize policy learning within the target environment.

Algorithm 1 Conceptual Overview of Counterfactually Guided Policy Transfer

```
for $k = 1, \dots, K$ do
    Sample counterfactual trajectories $\{\tau^i\}$ following $\pi^{(\mathtt{T})}$, using $P^{(\mathtt{S})}$ as a prior    ▷ See Algorithm 2
    Update $\pi^{(\mathtt{T})}$ using $\{\tau^i\}$ and Policy Iteration, regularized by $\pi^{(\mathtt{S})}$    ▷ See Algorithm 4
end for
```

subpopulations, with slightly differing dynamics, within each are unknown. These characteristics restrict accurate inference of the effect treatments have on sequential observations. As such, we facilitate guided learning in the target environment through an informative prior derived via counterfactual inference from the source environment.

We model both the source and target environment as Gumbel-Max SCMs to leverage counterfactual stability when transferring quantities from the environment $\mathtt{S}$. We assume an optimal policy $\pi^{(\mathtt{S})}$ has been learned with observational data from a source environment $\mathtt{S}$ and that the empirical transition matrix, $P^{(\mathtt{S})}$, is accessible. We demonstrate that learning in the environment $\mathtt{T}$ can be improved by 1) leveraging $P^{(\mathtt{S})}$ to improve counterfactual trajectory sampling compared to those obtained naively from $\mathtt{T}$ and by 2) regularizing the policy distribution in $\mathtt{T}$ by the distribution of $\pi^{(\mathtt{S})}$.

We first discuss how to estimate the *counterfactual* reward of executing a policy π in $\mathtt{T}$, using an empirical estimate of $P^{(\mathtt{T})}$ from trajectories generated by a behaviour policy $\mu^{(\mathtt{T})}$ and their corresponding outcome. This requires i) Estimating posteriors over exogenous variables $p(\mathbf{U}^{(\mathtt{T})}|\mathbf{X}^{(\mathtt{T})})$, ii) intervening with policy π in environment $\mathtt{T}$ with modified posterior estimates over exogenous variables, iii) collecting counterfactual trajectories, and iv) estimating the reward. Only step i) can be reliably executed purely from observational data. Without further assumptions on the underlying structure of the SCM, estimating counterfactual quantities like expected rewards under counterfactual roll-outs may be non-identifiable [46].

Counterfactual stability helps alleviate this challenge. If a discrete outcome i has been observed in the target environment, then under another intervention I' the likelihood of all other (counterfactual) outcomes $j \neq i$ is 0, or $P^{(\mathtt{T})do(\pi)|X_{I(\mu(T))}=i}(X=j) = 0$. We propose $p(\mathbf{g}^{(\mathtt{S})}|\mathbf{X}^{(\mathtt{S})})$ where $\mathbf{g}^{(\mathtt{S})} \subseteq \mathbf{U}^{(\mathtt{S})}$ as an informative prior from $\mathtt{S}$ on Gumbels in the target $p(\mathbf{g}^{(\mathtt{T})})$ while maintaining this stability. The benefits of regularization via an informative prior is two-fold: First, if the source and target environment are similar, using the prior improves transition estimation to benefit policy learning. Second, if environments differ the prior allows a form of guided exploration in data-scarce settings. We show this with a new policy learning method, Counterfactual Policy Iteration (Sec 3.3).

3.1 Improving counterfactual rollouts with an informative prior

We propose a mixture distribution over the Gumbels to sample from $p(\mathbf{g}^{(\mathtt{T})}|\cdot)$ as if parametrized by source transitions using the procedure outlined in Algorithm 2 instead of naively transferring transition dynamics from $\mathtt{S}$. In particular, when counterfactual trajectories are sampled without transfer, a posterior sampling of Gumbel variables is obtained via a procedure called Topdown

Algorithm 2 Modified Top-down with informative prior

```
   Procedure is repeated every step of a counterfactual roll-out to generate $\tau^i$
   Note: $\log P(s'|s,a) = \log \alpha$
1: procedure MIXTURE-TOPDOWN(SCM $\mathcal{M}$, source stats: $\log \alpha^{(\mathtt{S})}$, target stats: $\log \alpha^{(\mathtt{T})}$, counterfactual policy $\pi$ or $\log \hat{\alpha}^{(\mathtt{T})}$, mixture param $w^{\mathtt{T}}$, num of gumbel obs $N'$)
       // Gather a batch of counterfactual trajectories
2:     for $n' = 1, \ldots, N'$ do
3:         $\lambda \sim Bernoulli(w^{(\mathtt{T})})$                    ▷ Sample $\lambda$ to select mixture component
4:         $\log \alpha = \lambda \log \alpha^{\mathtt{T}} + (1-\lambda) \log \alpha^{\mathtt{S}}$           ▷ Fix Gumbel location
5:         $g_{cf}$ = Topdown($\log \alpha, 1$)                  ▷ Sample 1 or more posterior gumbels
6:         $s_{cf}^{n'} = \arg\max_j \log \hat{\alpha}^{(\mathtt{T})} + g_{cf}$             ▷ Sample counterfactual states
7:     end for
8:     $\hat{p}^{(\mathtt{T})}$ is the empirical estimate using $\{s_{cf}^{n'}\}_{n'=1}^{N'}$
9: end procedure
```

Gumbels [37] where the location estimates are fixed to target transition statistics. When leveraging the prior from $\mathtt{S}$, we sample from a mixture of Gumbel distributions determined by the (potentially noisy and biased) target and source transition estimates. A hyperparameter $w^{(T)}$ determines the amount of regularization i.e. proportion of times the samples come from either source vs target transitions. The sampling done in Algorithm 2 maintains counterfactual stability as the Gumbel mixture samples g_{cf} (Line 5) are fixed while sampling in Line 6 (see Appendix B for a formal justification). The resulting counterfactual rollouts help re-estimate transition dynamics in the target environment and reflect a form of exploration (see Algorithm 3 Line 5).

3.2 Regularized Policy Iteration

Given a fully observable finite and discrete MDP, Policy Iteration (PI) is guaranteed to converge to an optimal policy. PI is performed by switching between evaluation and improvement steps that estimate and refine a value function $V(s)$ and greedy policy π. However, this cycle may not fully converge if the MDP is partially observed, not accurately approximated or contains intractably large state and action spaces [61]. While convergence using PI is not guaranteed in our setting, the counterfactually sampled trajectories (Sec. 3.1) can help improve the accuracy of the transition matrix used in the evaluation step of PI. However, acting greedily with respect to the value function may still encourage poor behavior given the limited accuracy of $\tilde{P}^{(\mathtt{T})}$. We further seek to guide the development of the target policy by regularizing the policy improvement step of PI to not deviate too far from the source policy $\pi^{(\mathtt{S})}$. The intuition here is that $\pi^{(\mathtt{S})}$ has been exposed, in the data-rich $\mathtt{S}$, to a more accurate approximation of the true dynamics as well as transitions not in $\mathtt{T}$, positively influencing $\pi^{(\mathtt{T})}$.

We regularize PI through minimizing the KL-divergence between the policy distributions over actions, conditioned on the observed state. As we have limited our investigations to discrete and finite POMDPs this is equivalent to log-aggregation [24]. Within the policy improvement step we generate a proposal distribution $\nu(\cdot|s)$ over the actions:

$$\nu(\cdot|s) = \frac{1}{Z_p}\left(\mathcal{R}(s,\cdot) + \gamma\left(\tilde{P}^{(\mathtt{T})}(s'|s,\cdot) \otimes \mathbf{V}(s')\right)\right) \tag{1}$$

where Z_p is a normalization constant and the Kronecker operator $\otimes$ is used to denote a Matrix-vector product such that $V(s')$ is combined with $\tilde{P}^{(\mathtt{T})}(s'|s,\cdot)$ for each action and possible successor state s'. We then seek to find the policy that minimizes the divergence between $\nu(\cdot|s)$ and $\pi^{(\mathtt{S})}(\cdot|s)$. That is,

$$\pi_{k-1}^{(\mathtt{T})} = \arg\min_{\pi}\ \lambda\,\mathrm{KL}(\pi\|\nu) + (1-\lambda)\,\mathrm{KL}(\pi\|\pi^{(\mathtt{S})}) \tag{2}$$

where λ is set to limit how much $\pi^{(\mathtt{S})}$ influences $\pi^{(\mathtt{T})}$ and is selected empirically. See Section D for the derivation of Equation 2 and Algorithm 4 to see how it is implemented in REGPI.

3.3 Counterfactual Policy Iteration

The previous two subsections outline our major contributions toward the development of counterfactual policy iteration (CF-PI), the core method of our proposed CFPT procedure, visualized in Figure 1 and outlined in Algorithm 3 (full procedure found in Sec. 3.4). When learning a policy in the target

Algorithm 3 Counterfactual Policy Iteration

```
   // Counterfactual Policy Iteration (CF-PI)
1: procedure CF-PI( SCM $\mathcal{M}$, init. policy $\pi_0^{(\mathtt{T})}$, source policy $\pi^{(\mathtt{S})}$, source stats $P^{(\mathtt{S})}$, num. iters $K$, num. traj samples $N$, mixture params $\eta$ and $\lambda$)
2:     for $k = 1, \ldots, K$ do
           // Gather a batch of counterfactual trajectories
3:         $\{h^i\}_{i=1}^N \sim \mathcal{H}^{(\mathtt{T})} \subset \mathcal{T}$                                   ▷ Sample from observed data
4:         $\{\tau^i\}_{i=1}^N = \text{CFI}(\{h^i\}_{i=1}^N, \mathcal{M}, I(\mu \rightarrow \pi_{k-1}^{(\mathtt{T})}), \mathcal{T}, P^{(\mathtt{S})})$        ▷ CF rollouts under $\pi_{k-1}^{(\mathtt{T})}$
           // Estimate empirical transition statistics $\hat{P}^{(\mathtt{T})}$ from $\{\tau^i\}_{i=1}^N$
5:         $\tilde{P}^{(\mathtt{T})} = \frac{1}{Z_{\mathtt{T}}} \left( \eta \, P^{(\mathtt{T})} + (1-\eta) \, \hat{P}^{(\mathtt{T})} \right)$                  ▷ Augment transition stats
           // Regularized policy iteration with CF augmented transition stats
6:         $\pi_k^{(\mathtt{T})} \leftarrow \text{RegPI}(\pi_{k-1}^{(\mathtt{T})}, \gamma, \tilde{P}^{(\mathtt{T})}, \pi^{(\mathtt{S})}, \lambda)$
7:     end for
8: end procedure
```

environment $\mathtt{T}$, where a limited number of trajectories $\mathcal{H}^{(\mathtt{T})}$ have been collected following a behavior policy μ, we assume access to an optimal policy distribution $\pi^{(\mathtt{S})}$ as well as transition statistics $P^{(\mathtt{S})}$ from a relevant source environment $\mathtt{S}$. In clinical practice, $P^{(\mathtt{S})}$ may correspond to a communicated patient physiological response to a prescribed treatment while $\pi^{(\mathtt{S})}$ may reflect accepted treatment decisions when presented with a particular patient context. Policy learning via CFPT, with these elements from the source environment $\mathtt{S}$, is carried out through a counterfactually augmented form of Policy Iteration (CF-PI) taking CF-GPS [7] as inspiration.

CF-PI consists of K iterations where, in each iteration, a batch of counterfactual trajectories $\{\tau^i\}$ from $\mathtt{T}$ (Sec. 3.1)—sampled according to the current policy $\pi_{k-1}^{(\mathtt{T})}$—are used to augment the transition statistics $P^{(\mathtt{T})}$ for use in a regularized Policy Iteration subprocess (RegPI, Sec. 3.2) to update the policy $\pi^{(\mathtt{T})}$. This augmentation (Alg. 3, line 5) is a re-normalized weighted sum between the observed $P^{(\mathtt{T})}$ and the estimated transition statistics $\hat{P}^{(\mathtt{T})}$ drawn from $\{\tau^i\}$. The weighting parameter η is empirically chosen (see Section E.2.1 in the Appendix) to heavily favor observed transition statistics while still incorporating added diversity through counterfactual sampling, where Z_T is the normalizing constant over all successor states $S' = s'$ from any given state s. RegPI iterates between evaluation and improvement steps, where the source policy $\pi^{(\mathtt{S})}$ is used to regularize the improvement step. RegPI is run until convergence or for a maximum number of iterations.

3.4 CFPT Procedure

Here we present the psuedocode (Algorithm 4) outlining our proposed Counterfactually Guided Policy Transfer (CFPT) approach as discussed in this section. CFPT is enabled by first having access to an optimal treatment policy $\pi^{(\mathtt{S})}$ developed within a data-rich source environment $\mathtt{S}$ as well as an estimation of the transition statistics $P^{(\mathtt{S})}$ collected from observed data. These methods combine to form a two-phase counterfactual regularization approach for policy learning in a data-scarce target environment $\mathtt{T}$.

Policy learning is done through a counterfactually regularized form of PI (CF-PI). The heart of CF-PI rests on the discussion provided in Section 3.2 which introduces how we regularize PI (RegPI) in the target environment through KL-divergence log aggregation. CF-PI is executed as follows. For K iterations, a batch of trajectories $\{h^i\}_{i=1}^N$ observed within the target environment are sampled (Alg. 4, line 24). This batch is used, along with the current policy within $\mathtt{T}$, $\pi_k^{(\mathtt{T})}$, and the prior over the transition statistics from the source environment $P^{(\mathtt{S})}$ to generate counterfactual trajectories $\{\tau^i\}_{i=1}^N$ (Alg. 4, line 25 → CFI lines 1-6). This counterfactual sampling procedure, leveraging the property of counterfactual stability within Gumbel-Max SCMs, is described in Sections 3.1 and B (Appendix). The batch of trajectories produced may exhibit some diversity in observed transition statistics from those observed in $\mathtt{T}$. To account for this, an augmented transition matrix $\tilde{P}^{(\mathtt{T})}$ is formed through a weighted sum between $P^{(\mathtt{T})}$ and the empirically observed set from $\{\tau^i\}_{i=1}^N$ ($\hat{P}^{(\mathtt{T})}$, line 26). This

Algorithm 4 Counterfactually Guided Policy Transfer

```
    // Counterfactual inference (CFI) with source environment prior
1:  procedure CFI(data $\hat{x}_o$, SCM $\mathcal{M}$, intervention $I$, query $X_q$, prior $X_P^{(\mathrm{S})}$)
2:      $\hat{u} \sim p(u|\hat{x}_o)$      ▷ Sample noise variables from posterior over latent parameters
3:      $p(u) \leftarrow \delta(u - \hat{u})$      ▷ Replace noise distribution in $p$ with $\hat{u}$
4:      $f_i \leftarrow f_i^I$      ▷ Perform intervention $I$
5:      return $x_q \sim p^{\mathrm{do}(I)}(x_q|\hat{u})$      ▷ Simulate from the counterfactual posterior over model $\mathcal{M}^I_{\hat{x}_o}$, Alg. 2
6:  end procedure
    // Regularized Policy Iteration (RegPI)
7:  procedure REGPI(current policy $\pi^{(\mathrm{T})}$, discount $\gamma$, aug. statistics $\tilde{P}^{(\mathrm{T})}$, source policy $\pi^{(\mathrm{S})}$, reg. param $\lambda$)
8:      Initialize $V(s)$ for all $s \in \mathcal{S}$
9:      repeat
10:         repeat      ▷ // Policy Evaluation
11:             for each $s \in \mathcal{S}$ do
12:                 $v \leftarrow V(s)$
13:                 $V(s) \leftarrow \sum_{s'} \tilde{P}^{(\mathrm{T})}(s'|s, \pi^{(\mathrm{T})}(s)) \left[\mathcal{R}(s, \pi^{(\mathrm{T})}(s)) + \gamma V(s')\right]$
14:             end for
15:         until convergence
16:         for each $s \in \mathcal{S}$ do      ▷ // Policy Improvement
17:             $\nu(\cdot|s) \leftarrow \frac{1}{Z_p}\left(\mathcal{R}(s,\cdot) + \gamma\left(\tilde{P}^{(\mathrm{T})}(s'|s,\cdot) \otimes \mathbf{V}(s')\right)\right)$      ▷ Gen. a proposal dist. over actions
18:             $\pi^{(\mathrm{T})}(s) \leftarrow \arg\max_a \exp\left\{\lambda \log \nu(a|s) + (1-\lambda)\log \pi^{(\mathrm{S})}(a|s)\right\}$ ▷ KL minimization, Eq. 19
19:         end for
20:     until $\pi^{(\mathrm{T})}$ converges or after MAX_ITERATIONS
21: end procedure
    // Counterfactual Policy Iteration (CF-PI)
22: procedure CF-PI( SCM $\mathcal{M}$, init. policy $\pi_0^{(\mathrm{T})}$, source policy $\pi^{(\mathrm{S})}$, source statistics $P^{(\mathrm{S})}$, num. iters $K$,
    num. traj samples $N$, mixture param $\eta$)
23:     for $k = 1, \ldots, K$ do
        // Gather a batch of counterfactually generated trajectories in the target environment
24:         $\{h^i\}_{i=1}^N \sim \mathcal{H}^{(\mathrm{T})} \subset \mathcal{T}$      ▷ Sample batch of trajectories from observed data
25:         $\{\tau^i\}_{i=1}^N = \mathrm{CFI}(\{h^i\}_{i=1}^N, \mathcal{M}, I(\mu \to \pi_{k-1}^{(\mathrm{T})}), \mathcal{T}, P^{(\mathrm{S})})$      ▷ Counterfactual rollouts under $\pi_{k-1}^{(\mathrm{T})}$
        // Estimate empirical transition statistics $\hat{P}^{(\mathrm{T})}$ from $\{\tau^i\}_{i=1}^N$
26:         $\tilde{P}^{(\mathrm{T})} = \frac{1}{Z_{\mathrm{T}}}\left(\eta\, P^{(\mathrm{T})} + (1-\eta)\, \hat{P}^{(\mathrm{T})}\right)$      ▷ Augment observed environment transition statistics
        // Regularized policy iteration with counterfactually augmented target env. transition statistics
27:         $\pi_k^{(\mathrm{T})} \leftarrow \mathrm{RegPI}(\pi_{k-1}^{(\mathrm{T})}, \gamma, \tilde{P}^{(\mathrm{T})}, \pi^{(\mathrm{S})}, \lambda)$
28:     end for
29: end procedure
```

augmented transition matrix is then passed to RegPI as discussed in Section 3.2 (line 27 → RegPI, lines 7-21).

RegPI alternates between policy evaluation and policy improvement steps. In policy evaluation (lines 11-15) where the current policy $\pi_k^{(\mathrm{T})}$ is used to refine an estimate of the underlying value function based on the observed rewards and estimated transition statistics when applying $\pi_k^{(\mathrm{T})}$. Once this value estimate converges, it is used in a form of a Bellman update (line 17) to generate a proposal distribution over actions for each state. This is the beginning of the policy improvement step (lines 16-19). After the proposal distribution $\nu(\cdot|s)$ is generated, it is used to estimate the best policy while being constrained by the source policy $\pi^{(\mathrm{S})}$ through KL-divergence log-aggregation (line 18). This improved policy is then sent back to the evaluation step to refine the estimate of the value function and this process continues until $\pi_k^{(\mathrm{T})}$ converges or a maximum number of iterations has been performed. With this updated policy, a new batch of trajectories are sampled from $\mathcal{H}^{(\mathrm{T})}$ to draw new counterfactual samples and next iteration continues to further optimize the target policy $\pi^{(T)}$.

4 Experiments

We demonstrate the benefits of CFPT through a coarse simulation of prescribing treatment to septic patients. This simulator has been used to demonstrate that off-policy evaluation with limited observational data in partially observed settings can be misleading [17, 43][1]. The simulator approximates patient physiology (discretized vital measurements of heart rate, blood pressure, oxygen concentration, and glucose levels) in response to medical interventions and a latent state representing whether the patient is diabetic. Potential treatments include binary selections of antibiotics, vasopressors, and mechanical ventilation. Rewards are given when a patient is discharged (+1) or dies (-1). Discharge occurs after all vitals are 'normal' and all treatments discontinued; death occurs if any three of the vitals are out of 'normal' range simultaneously. We simulate domain shift in the simulator by varying the proportions of diabetic patients between the source and target environments. Hidden confounding is induced by masking the patient's diabetic state which makes the transfer task more challenging. Our objective is to demonstrate that, through CFPT, we can improve the baseline policy performance when transferring to a more challenging and disparate target environment where the proportion of diabetic patients is much higher. Additionally we leverage counterfactual off-policy evaluation (CF-PE) for introspection of our learned policies to illustrate where a policy learned through CFPT improves over baseline policies[2].

Baselines: To illustrate the benefits of CFPT in a data-limited target environment T, we compare to several baselines and ablations. i) **RANDOM** where actions are chosen randomly, ii) **SCRATCH** policy learned using policy iteration (PI) natively using observational data $\mathcal{H}^{\texttt{T}}$. If there are any observable benefits from transfer, we expect to easily outperform these two baselines. The next two baselines are naive transfer approaches leveraging separate aspects of the source environment S. iii) **POOLED** pools observed data between environments to learn a policy in T using PI, and iv) **BLIND** applies the source policy $\pi^{(\texttt{S})}$ in T without any adaptation.. Each of these practices are infeasible in most clinical settings due to privacy and safety concerns, yet we include them for completeness. We also compare CFPT to two ablations showcasing the benefits of each contribution outlined in Section 3. v) **REGPI** omits counterfactual batch sampling, only regularizing $\pi^{(\texttt{T})}$ by its divergence from $\pi^{(\texttt{S})}$ within policy iteration (cf. Sec. 3.2). This, in essence, is a form of fine-tuning $\pi^{(\texttt{S})}$ in T with the parameter λ representing how aggressive the adaptation is. vi) **RED. CFPT** is a reduced form of CFPT where we omit the informative prior from S over the transition statistics when sampling counterfactual trajectories. Here, the counterfactual trajectories are drawn according to the Gumbel variables from T and policy learning is done through RegPI . All parameter settings used to train these policies are included in Section E of the Appendix.

Setup: The behavior policy μ was found via PI with full access to the MDP (including diabetes state) to provide a strong observation policy, following [43]. When generating the observed trajectories $\mathcal{H}$, the policy takes random actions w.p. 0.15 to introduce some variation. Within S, $|\mathcal{H}^{(\texttt{S})}| = 10,000$, with at most 20 steps per trajectory. Additionally, the probability of a trajectory being from a diabetic patient (a stochastic Glucose response) in S is 0.1, where the SCRATCH baseline performs best (see Fig 2a). We set T to be substantially harder to learn in, limiting $\mathcal{H}^{(\texttt{T})}$ to contain only 2000 trajectories, with the primary target environment having a diabetic subpopulation proportion set to 0.8 (where SCRATCH performs worst). Given the prevalence of diabetic patients and the limited number of trajectories, we anticipate the transition statistics $P^{(\texttt{T})}$ will be far from true dynamics governing both S and T, providing opportunity to demonstrate the benefits of careful transfer from S.

4.1 Results

4.1.1 Quantitative Evaluation of $\pi^{(\texttt{T})}$

CFPT under domain shift: Figure 2(a) demonstrates the average reward of all baselines and CFPT as calculated by CF-PE (dashed lines) and applying the learned $\pi^{(\texttt{T})}$ to simulate an additional 5000 trajectories in T (solid lines) as we change proportions of diabetic patients in the target using a fixed source environment. The benefits of CFPT (blue) are clear across all levels of domain shift, with significant benefits when mixture populations in the target are skewed. Diabetic patients are harder to treat, resulting in a general decreasing trend in true reward as the proportion of diabetic patients increases. For CFPT, the advantages of leveraging the source policy in a distributionally

[1] `https://github.com/clinicalml/gumbel-max-scm`

[2] Our code will be made available in the near future

similar target environment (pDiab= 0.3) is clear while in disparate environments CFPT provides guided perturbations through counterfactual inference to improve policy learning. The clearest advantage of CFPT is that it settles on and improves over baselines as the proportions of patient subpopulations within T invert with respect to S. CFPT also reflects less over-confidence according to CF-PE (dashed lines) among target populations with an imbalanced composition of patient types. This communicates that the policy has learned more robust and accurate representations of observation-treatment dynamics with respect to expected outcomes, matching with true policy performance more closely. Overall, this quantitative evaluation is a strong indication of the benefits of our proposed two-fold regularization when faced with domain shift between environments.

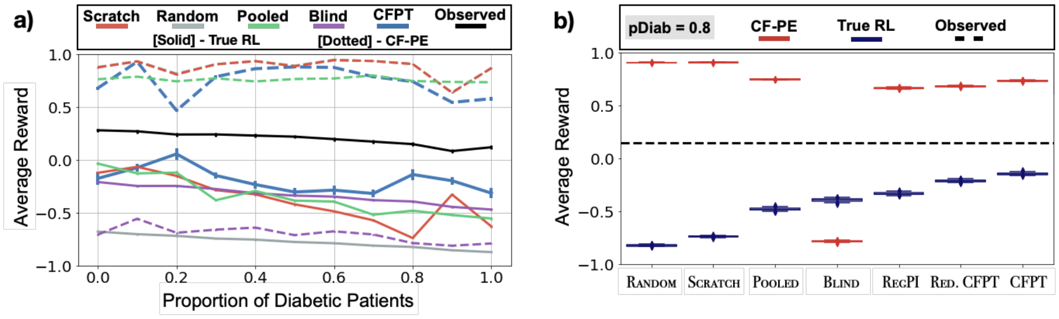

Figure 2: Quantitative comparison between CFPT and the baselines proposed in Sec. 4 based on estimated reward from the learned policy in T. Results after counterfactual policy evaluation (CF-PE) are plotted in red where the true performance in T is plotted in blue. 95% uncertainty intervals are found through 100 bootstrapped samples of the 5000 generated trajectories under the learned target policy. CFPT outperforms all comparisons and most closely approximates the observed behavior.

Ablation Study for CFPT: In Figure 2(b) we present the average reward for all baselines and CFPT as calculated by CF-PE (in red) and true rewards (in blue) for our primary target environment (0.8 proportion of diabetic patients). As expected, CF-PE overestimates the true RL return in T, even with pathologically poor policies (i.e. Random). Of note however is how the policies learned under transfer perform in T. The baselines Pooled and Blind provide significant improvements over the naively learned Scratch baseline. However with each additional component (RegPI → Red. CFPT → CFPT) of our proposed procedure, the true RL return monotonically improves over each baseline and approaches the observed return of the behavior policy the data was generated from. It is also notable that under the variants of CFPT, the off-policy evaluation of the learned $\pi^{(\mathtt{T})}$ appears to experience a reduction in overestimating the performance. As we demonstrate through counterfactual introspection in the following experiment, the policy learned through CFPT acts circumspectly and closely approximates the observed behavior, leading to more stable performance.

4.1.2 Qualitative Evaluation of $\pi^{(\mathtt{T})}$

Treatment Selection under CFPT: To better compare policy evaluations between baselines, we perform an introspective analysis using CF-PE on both a policy and trajectory level. First, we compare the counterfactual outcomes between the naive baseline policy without transfer (Scratch) against our full CFPT trained policy, to identify how CFPT improves policy learning within T (other comparisons between CFPT and the baselines are in Section E.2 in the Appendix). We first compare the counterfactual outcomes as estimated through CF-PE and then compare policy behavior under counterfactual evaluation for an individual patient drawn from T. In Figure 3 (left) we present the aggregate counterfactual outcomes as suggested by CF-PE in comparison to what was observed. The primary difference in the evaluation of these two policies is in the percentage of patients CFPT does not discharge while Scratch does. To further identify what separates these two policies we select patients who die under the behavior policy but are inferred to be discharged under Scratch but kept in the hospital under CFPT. In Figure 3 (right), we observe that the non-transfer baseline (Scratch) is far more aggressive in it's treatment decisions, leading to premature treatment cessation as the patient's condition deteriorates (visualized by the blue counterfactual trajectories) immediately after they are indicated for discharge. In contrast, the CFPT policy chooses a strategy that stably maintains the patient condition, continuing all treatments until the observation window terminates.

Subpopulation Analysis: A subpopulation level analysis (diabetic vs non-diabetic) under CF-PE for both CFPT and Sratch is included in Figure 4 and continued in Section E.2.2 in the

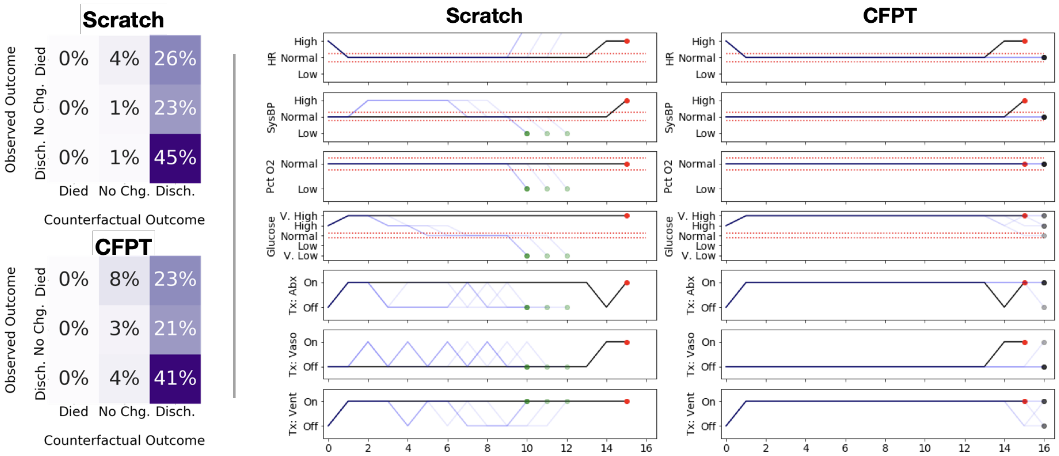

Figure 3: Qualitative comparison between CFPT (Scratch). At left, we compare the aggregate statistics of CF-PE suggested outcomes against the true observed outcome. At right, we compare an individual patient's counterfactual trajectories using these policies. Dark lines are the observed vital measurements and actions over time while the lighter blue traces correspond to counterfactual observations and actions. Green, red and black markers denote discharge, death and no change respectively. CFPT provides more stable treatment selection in comparison with the non-transfer baseline. Additional samples in Appendix E.2

.

Appendix. CFPT, in aggregate, demonstrates a more measured approach compared to Scratch, as there isn't as strong of consensus toward discharge. Additional subpopulation analysis for different levels of regularization as well as different proportions of target diabetic populations is provided in Appendix E.2. Through this introspective analysis, we conclude that CFPT has developed a more sustainable and reliable policy through transfer. In connection with the quantitative evaluation, it is clear that CFPT produces better policies and that the counterfactually guided transfer approach we have proposed in this paper is promising for learning in data-limited off-policy settings.

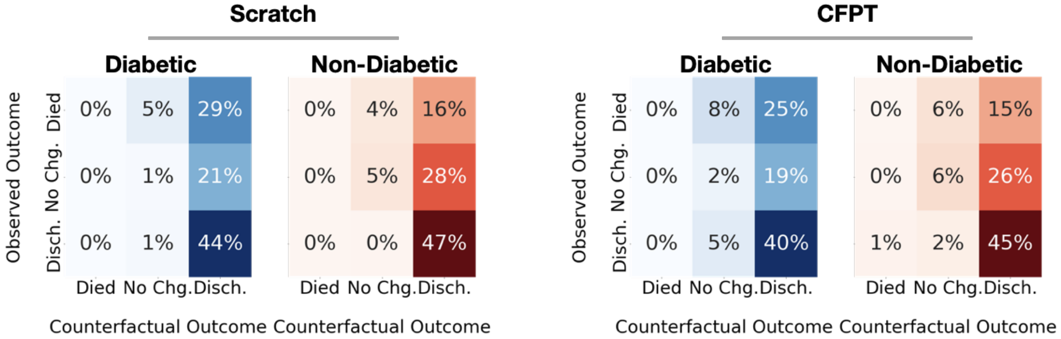

Figure 4: Aggregated counterfactual outcomes by subpopulation following the non-transfer baseline policy vs CFPT. These values are normalized by the number of patients belonging to each subpopulation (diabetic vs. non-diabetic) respectively. CFPT in aggregate is more conservative for the diabetic (rare class in source) in CF-PE evaluation.

5 Discussion and Conclusion

Motivated by challenges of computational knowledge transfer between clinical settings, we have introduced Counterfactually Guided Policy Transfer. This procedure builds from counterfactual stability within Gumbel-Max SCMs to leverage elements of a data-rich clinical setting to facilitate better learning in a data-scarce setting. We utilize two components of the source environment in our transfer framework: 1) The observed transition statistics $P^{(\mathrm{S})}$ and 2) the trained treatment policy $\pi^{(\mathrm{S})}$. In clinical practice, $P^{(\mathrm{S})}$ may correspond to a communicated aggregate patient physiological response to a prescribed treatment while $\pi^{(\mathrm{S})}$ reflects accepted treatment decisions when presented a particular patient context. Both these elements can be feasibly shared in a secure and private manner and, as demonstrated by this work, used to improve treatment policy development in a disparate clinical setting.

The work we have presented in this paper stands as an initial step in the development of counterfactually-aided policy transfer to reliably extend model usage beyond the environment it was trained in. While the discrete setting we have used in this work is suitable for an intriguing, theoretically grounded, proof of concept it is far from representative of an actual clinical setting. The limited empirical study used to illustrate and further demonstrate the methodological contributions we introduce in this paper is not sufficient to support any immediate implementation in an actual clinical setting. To do so would be highly unethical, reckless and would put patient lives at risk.

At present, the mixture of the transition statistics and the policy regularization is empirically fixed. In future work, we endeavor to adjust the level of mixture and regularization adaptively based on the uncertainty of the transition statistics and treatment selection process, respectively. The work we have presented in this paper stands as an initial step in the development of counterfactually-aided policy transfer to reliably extend learned models beyond the environment it was trained in. While the discrete setting we have used in this work is suitable for a proof of concept, we intend to broaden the theoretical foundation supporting our procedure to admit continuous observations and treatments to potentially support policy development in true-to-life simulated settings and eventually with retrospective medical data.

Continued development, building from transfer approach we have proposed in this work, will need to assure that any transferred model produces equitable benefit across demographic subgroups. Possible sources of hidden confounding and physician bias present in recorded data may be derived from patient age, gender, race, socio-economic status among other factors [8, 42, 66]. While remaining cognizant of these important factors, enforcing fairness and reducing bias within our trained treatment policies, we are highly motivated by the promise of developing a paradigm that allows for reliable and robust computational knowledge transfer between medical institutions. This advancement, if successful, would have the potential to offer state-of-the-art treatment policies in recognition of individual patient conditions to small regional hospitals that would not have the means to develop their own machine learning-based clinical decision support tools due to data limitations and computational resources. In short, the technological and medical advances being made in large research hospitals through machine learning would be available broadly as conditions between the hospitals allow.

Acknowledgments and Disclosure of Funding

We thank our many colleagues who contributed to several thoughtful discussions and provided timely advice to improve this work. We specifically appreciate the feedback provided by Sindhu Gowda, Nathan Ng, Elliot Creager, David Madras and Vinith Suriyakumar.

This research was supported in part by Microsoft Research, a CIFAR AI Chair at the Vector Institute, a Canada Research Council Chair, and an NSERC Discovery Grant

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute `www.vectorinstitute.ai/#partners`.

References

[1] Abbas Abdolmaleki, Jost Tobias Springenberg, Jonas Degrave, Steven Bohez, Yuval Tassa, Dan Belov, Nicolas Heess, and Martin Riedmiller. Relative entropy regularized policy iteration. *arXiv preprint arXiv:1812.02256*, 2018.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 5–6, 2015.

[4] Xiaowu Bai, Wenkui Yu, Wu Ji, Zhiliang Lin, Shanjun Tan, Kaipeng Duan, Yi Dong, Lin Xu, and Ning Li. Early versus delayed administration of norepinephrine in patients with septic shock. *Critical care*, 18(5): 532, 2014.

[5] Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. In *Advances in neural information processing systems*, pages 280–288, 2014.

[6] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pages 4055–4065, 2017.

[7] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

[8] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.

[9] Christoph Dann, Gerhard Neumann, Jan Peters, et al. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

[10] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516, 2019.

[11] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

[12] Amir M Farahmand, Mohammad Ghavamzadeh, Shie Mannor, and Csaba Szepesvári. Regularized policy iteration. In *Advances in Neural Information Processing Systems*, pages 441–448, 2009.

[13] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1): 4809–4874, 2016.

[14] Mahdi Milani Fard, Joelle Pineau, and Peng Sun. A variance analysis for POMDP policy evaluation. In *AAAI*, pages 1056–1061, 2008.

[15] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[16] Vincent François-Lavet, Guillaume Rabusseau, Joelle Pineau, Damien Ernst, and Raphael Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *Journal of Artificial Intelligence Research*, 65:1–30, 2019.

[17] Joseph Futoma, Michael C Hughes, and Finale Doshi-Velez. Popcorn: Partially observed prediction constrained reinforcement learning. *arXiv preprint arXiv:2001.04032*, 2020.

[18] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

[19] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1 (4):e157–e159, 2019.

[20] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25(1):16–18, 2019.

[21] Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pages 2366–2375, 2019.

[22] Hado V Hasselt. Double q-learning. In *Advances in neural information processing systems*, pages 2613–2621, 2010.

[23] Tamir Hazan and Tommi Jaakkola. On the partition function and random maximum a-posteriori perturbations. *arXiv preprint arXiv:1206.6410*, 2012.

[24] Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In *Advances in neural information processing systems*, pages 266–272, 1998.

[25] Matteo Hessel, Hado van Hasselt, Joseph Modayil, and David Silver. On inductive biases in deep reinforcement learning. *arXiv preprint arXiv:1907.02908*, 2019.

[26] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

[27] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[28] Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8895–8906, 2018.

[29] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in neural information processing systems*, pages 9269–9279, 2018.

[30] Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. In *Advances in neural information processing systems*, pages 6250–6261, 2017.

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

[32] Romain Laroche and Merwan Barlier. Transfer reinforcement learning with shared dynamics. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[33] Alessandro Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.

[34] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578, 2018.

[35] Sanghack Lee, Juan D Correa, and Elias Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains. Technical report, Technical Report R-52, Causal AI Lab, Department of Computer Science . . . , 2020.

[36] RD Luce. Individual choice behavior: A theoretical analysis: Courier corporation. 2005.

[37] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems*, pages 3086–3094, 2014.

[38] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[39] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance in value function estimation. In *Proceedings of the twenty-first international conference on Machine learning*, page 72, 2004.

[40] Paul E Marik, Walter T Linde-Zwirble, Edward A Bittner, Jennifer Sahatjian, and Douglas Hansell. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. *Intensive care medicine*, 43(5):625–632, 2017.

[41] Vukosi Ntsakisi Marivate, Jessica Chemali, Emma Brunskill, and Michael Littman. Quantifying uncertainty in batch personalized sequential decision making. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[42] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[43] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890, 2019.

[44] Trishan Panch, Heather Mattie, and Leo Anthony Celi. The "inconvenient truth" about ai in healthcare. *Npj Digital Medicine*, 2(1):1–3, 2019.

[45] Sonali Parbhoo, Mario Wieser, and Volker Roth. Cause-effect deep information bottleneck for incomplete covariates. *arXiv preprint arXiv:1807.02326*, 2018.

[46] Judea Pearl. *Causality*. Cambridge university press, 2009.

[47] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[48] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

[49] Aniruddh Raghu, Omer Gottesman, Yao Liu, Matthieu Komorowski, Aldo Faisal, Finale Doshi-Velez, and Emma Brunskill. Behaviour policy estimation in off-policy policy evaluation: Calibration matters. *arXiv preprint arXiv:1807.01066*, 2018.

[50] Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.

[51] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[52] Joshua Romoff, Peter Henderson, Alexandre Piche, Vincent Francois-Lavet, and Joelle Pineau. Reward estimation for variance reduction in deep reinforcement learning. In *Conference on Robot Learning*, pages 674–699, 2018.

[53] Peter Schulam and Suchi Saria. Integrative analysis using coupled latent variable models for individualizing prognoses. *The Journal of Machine Learning Research*, 17(1):8244–8278, 2016.

[54] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017.

[55] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[56] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[57] Tavpritesh Sethi, Anant Mittal, Shubham Maheshwari, and Samarth Chugh. Learning to address health inequality in the united states with a bayesian decision network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 710–717, 2019.

[58] Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1-2):109–136, 2011.

[59] Adarsh Subbaswamy and Suchi Saria. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 947–957. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

[60] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 2020.

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

[62] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

[63] Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. In *Thirty-fourth AAAI conference on artificial intelligence*, 2020.

[64] Andrea Tirinzoni, Mattia Salvini, and Marcello Restelli. Transfer of samples in policy search via multiple importance sampling. In *International Conference on Machine Learning*, pages 6264–6274, 2019.

[65] Jason Waechter, Anand Kumar, Stephen E Lapinsky, John Marshall, Peter Dodek, Yaseen Arabi, Joseph E Parrillo, R Phillip Dellinger, and Allan Garland. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. *Critical care medicine*, 42(10):2158–2168, 2014.

[66] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[67] Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: a survey. *arXiv preprint arXiv:1908.08796*, 2019.

A Background: SCM

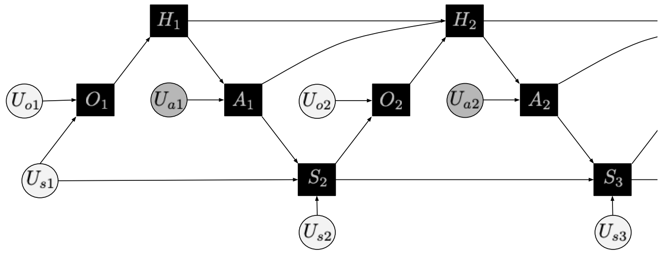

Figure 5: SCM of a POMDP Buesing et al. [7]. While nodes denote unobserved variables, gray nodes denoted observed latent variables and calculated quantitities are represented as black notes. Exogenous variables $\mathbf{U}$ are not affected by interventions. We assume this structure for both source and target environments.

A.1 Structural Causal Models [46]

A structural causal model $\mathcal{M}$ describes the causal mechanisms driving a system. It consists of an ordered triple $\langle \mathbf{U}, \mathbf{X}, \mathbf{F} \rangle$; a set of independent exogenous random variables $\mathbf{U} = \{U_1, U_2, \ldots, U_k\}$ that represent factors of variation outside the model, $\mathbf{X}$ comprises the endogenous variables modeled in the causal system and, the set of functions $\mathbf{F}$ defined by $X_i := f_i(\mathbf{PA}_i, U_i)\ \forall i$ where $\mathbf{PA}_i \subseteq \mathbf{X} \setminus X_i$ govern the causal mechanisms. $\mathbf{PA}_i$ are the parents of X_i in a causal DAG $\mathcal{G}$. The framework attributes probabilistic Markov assumptions to the joint distribution $P^{\mathcal{M}}$associated with the variables $(\mathbf{X}, \mathbf{U})$ in the graph. This characterizes a probability distribution, implying that one can observe samples true to the underlying causal graph and mechanism.

Definition A.1. *Interventional Distribution: An intervention I in an SCM $\mathcal{M}$ consists of replacing some functions $f_i(\boldsymbol{PA}_i, U_i)$ with a different governing causal mechanism $f_i^I(\boldsymbol{PA}_i^I, U_i)$ where $\boldsymbol{PA}_i^I$ are the parents of X_i in a new DAG $\mathcal{G}^I$. Note that the interventional distribution does not change the exogenous mechanisms driving the system. The resulting SCM, denoted by $\mathcal{M}^{do(I)}$ has a new joint distribution denoted by $P^{\mathcal{M}^{do(I)}}$.*

An intervention I is generally used to evaluate the *prospective* effect of perturbing the underlying causal mechanism. A more useful quantity in off-policy learning is the *counterfactual* which allows you to answer the causal queries of the form: "what would have happened had we given the patient medication b having observed no improvement with medication a?" Answering such *retrospective* queries requires inferring a model of the exogenous variables $P(\mathbf{U}|\mathbf{X} = \mathbf{x})$ and intervene with I on a causal system with exogenous noise priors $p(\mathbf{U})$ replaced by $p(\mathbf{U}|\mathbf{X} = \mathbf{x})$.

Definition A.2. *Counterfactual Distribution: Let $\mathcal{M}^{\mathbf{x}}$ correspond to the SCM where the exogenous noise model $p(\mathbf{U})$ in $\mathcal{M}$ is replaced by $p(\mathbf{U}|\mathbf{X} = \mathbf{x})$. Intervening with I on the resulting SCM $\mathcal{M}^{\mathbf{x}}$ yields a new SCM $\mathcal{M}^{do(I)|\mathbf{x}}$ and induces the joint counterfactual distribution $P^{\mathcal{M}^{do(I)|\mathbf{X}=\mathbf{x}}}$.*

A.2 Gumbel-Max SCM [43]

Definition A.3. *Gumbel-Max Trick: a sampling procedure from any discrete distribution with k categories, parametrized by $p_i = P(X = i), \forall i \in \{1, 2, \ldots, k\}$. First, sample k independent Gumbel variables g_j with location 0, scale 1. Set the sampled outcome $k = \arg\max_j \log p_j + g_j$.*

A Gumbel-Max SCM is one in which all nodes $\mathbf{X}$ are discrete random variables. Given independent Gumbel variables $\mathbf{g} = \{g_1, g_2, \ldots, g_k\}$, the causal mechanisms are given by: $X_i := f_i(\mathbf{PA}_i, g_i) = \arg\max_j \log p(X_i = j|\mathbf{PA}_i) + g_j$.

Non-identifiability of causal effect estimation under counterfactual scenarios is challenging for reliable transfer. That is, there may be multiple SCMs consistent with observations that provide different counterfactual estimates. In order to reliably draw causal conclusions from a counterfactual query, which is what we will need, further assumptions are required. In the case of binary SCMs, this assumption is given by the monotonicity condition [46] and in the discrete case known as counterfactual stability.

Let $P^{do(I)}(Y = i) = p_i\ \forall i \in [L]$ and $P^{do(I')}(Y = i) = p'_i\ \forall i \in [K]$. Let $P^{do(I)}(X = i)$ be the probability of observing i under intervention I for variable X in a discrete SCM and the observed outcome be represented by X_I. Then $P^{do(I')|X_I = i}(X = j)$ is the counterfactual probability of observing outcome j having observed i under intervention I.

Definition A.4. *Counterfactual Stability: An SCM over discrete random variables is counterfactually stable if: If we observe $X_I = i$, then $\forall j \neq i$, if $\frac{p_i'}{p_i} \geq \frac{p_j'}{p_j}$, implies that $P^{do(I')|X_I=i}(X = j) = 0$.*

B Informative mixture-prior for counterfactual regularization

Let $P^{(\mathtt{S})}$ be the estimated transition matrix in the source domain $\mathtt{S}$. Then we consider the potential prior distribution of the Gumbel exogenous variables in the target domain given by $\log P^{(\mathtt{S})}$. We further impose independent priors over each state-action pair in the target domain. That is, consider the following potential prior on Gumbel variables corresponding to some state-action pair s, a: $p_{s,a}(\mathbf{g}^{(\mathtt{T})}) = \prod_{i=1}^{K} f_{\log P^{(\mathtt{S})}(s'=i|s,a)}(g_i)$, where $f_{\log \alpha}$ is the density of a Gumbel random variable with location $\log \alpha$ and scale 1:

$$\begin{aligned} f_{\log \alpha}(g_i) &= \exp\left(-g + \log \alpha\right) \exp\left(-\exp\left(-g + \log \alpha\right)\right) \\ &= \exp\left(-g + \log \alpha\right) F_{\log \alpha}(g) \end{aligned} \tag{3}$$

where K is the number of discrete states and $F_{\log \alpha}(g)$ is the CDF of the Gumbel variable with location $\log \alpha$. Without an informative prior, these gumbels would be sampled with location 0 independently, instead of $\log P^{(\mathtt{S})}$.

In order to improve policy iteration, we hope to sample counterfactual trajectories by rolling out the current estimated learned policy in the target domain, using observational data collected from behaviour policy $\mu^{(\mathtt{T})}$. Let the current observation in the target domain be k'. Now consider the joint distribution of k' and $\mathbf{g}^{(\mathtt{S})}$ for any fixed state-action pair (we drop explicit notation for clarity). To account for the informative prior, we treat the locations of these Gumbel variables to be random-variables $\boldsymbol{\alpha}$. Thus to obtain the joint distribution, we integrate over $\boldsymbol{\alpha}$:

$$\begin{aligned} p(k', g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})}) &= \int_{\alpha} p(k', g_1^{(\mathtt{T})}, \ldots, g_K^{(\mathtt{T})}, , \alpha_1, \ldots, \alpha_n) d\boldsymbol{\alpha} && (4) \\ &= \int_{\boldsymbol{\alpha}} p(k', g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) d\boldsymbol{\alpha} && (5) \\ &= \int_{\boldsymbol{\alpha}} \frac{\alpha_{k'}}{Z} f_{\log Z}(g_{k'}^{(\mathtt{T})}) \prod_{i \neq k'} \left[f_{\log \alpha_i}(g_i^{(\mathtt{T})}) \frac{[\![g_{k'}^{(\mathtt{T})} \geq g_i^{(\mathtt{T})}]\!]}{F_{\log \alpha_i}(g_{k'}^{(\mathtt{T})})} \right] p(\boldsymbol{\alpha}) d\boldsymbol{\alpha} && (6) \end{aligned}$$

Equation (6) can be obtained exactly following[3] Maddison et al. [37]. That is,for a fixed and known α, Maddison et al. [37] show that in the posterior, the gumbel corresponding to the observed outcome k', i.e. $g_{k'}^{(\mathtt{T})}$ is a Gumbel variable with location parameter $Z = \log \sum_{i=1}^{K} \alpha_i$, the max value k' and Gumbels are independent and the rest of the exogenous variables $g_i^{(\mathtt{T})} \forall i \neq k'$ are truncated Gumbel variables (truncated by the max gumbel value corresponding to k' (shown by the Iverson brackets above). Without any prior on the Gumbels, the location parameters in the target domain can simply be obtained according to the transition probabilities estimated naively from limited data. That is $p(\boldsymbol{\alpha}) = \delta(\log P^{\mathtt{T}}(\cdot|\cdot))$ where δ is the dirac delta distribution. In a data-scarce environment, the transition estimates are prone to error leading to a bias in value function estimates [39, 14] for policy iteration. In our algorithm these estimation errors affect the quality of counterfactual trajectories. Thus, instead we propose uncertainty via a new prior over $\alpha' s$ and consequently the posterior estimates of the Gumbel parameters as follows:

$$p(\boldsymbol{\alpha}) = \begin{cases} \log P^{\mathtt{T}}(\cdot|\cdot) & \text{w.p. } 0 \leq w^{(\mathtt{T})} \leq 1 \\ \log P^{\mathtt{S}}(\cdot|\cdot) & \text{w.p. } w^{(\mathtt{S})} = 1 - w^{(\mathtt{T})} \end{cases}$$

Consequently, this imposes a mixture distribution over the *posterior* Gumbel distribution conditioned on observation k', given by:

$$p(g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | K) = w^{(\mathtt{T})} p(g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | K) + w^{(\mathtt{S})} p(g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | K)$$

The mixture contributions can be treated as a hyper-parameter that determines the amount of regularization from the source domain. This motivates the Modified top-down sampling procedure outlined in Algorithm 2. For every counterfactual sample, we first select the mixture component with probability $[w^{(\mathtt{T})}, 1 - w^{(\mathtt{T})}]$. Note that while this prior in-turn induces a prior probability on the discrete distributions, the transition-estimates for policy learning are not directly regularized using this prior but via empirical estimates coming from the counterfactual trajectories within the same environment.

[3] `https://cmaddis.github.io/gumbel-machinery`

B.1 Mixture-prior preserves counterfactual stability

Definition B.1. *Counterfactual Stability: An SCM over discrete random variables is counterfactually stable if: If we observe $X_I = i$, then $\forall j \neq i$, if $\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j}$, implies that $P^{do(I')|X_I=i}(X = j) = 0$.*

Our proof is based on the insight that counterfactual stability is invariant to choice of prior so long as the gumbel samples are fixed across interventions. Our modified topdown sampling procedure ensures the same gumbel samples are used across interventions. Hence we preserve counterfactual stability even with regularization. For completeness, we include the contrapositive proof of Oberst and Sontag [43] here:

As denoted before, let $X_I^{(\mathtt{T})} = i$ (we drop `T` from superscript for random variables when context is clear) be the outcome observed under intervention (behaviour policy) in the target domain. The state observation i implies almost surely:

$$\log p_i + g_i^{(\mathtt{T})} > \log p_j + g_j^{(\mathtt{T})} \forall j \neq i \tag{7}$$

where $p_i := P^{(\mathtt{T})}(X = i)$ is short hand for the state-transition probabilities in the target domain induced using the Mixture-prior described above. To prove counterfactual stability, the contrapositive is proved i.e. $\forall j \neq i, P^{do(I')|X_I=i}(X = j) \neq 0 \implies \frac{p'_i}{p_i} < \frac{p'_j}{p_j}$.

To begin with, if $P^{do(I')|X_I^{(\mathtt{T})}=i}(X^{(\mathtt{T})} = j) \neq 0$ implies that there exist gumbel variables $g_i^{(\mathtt{T})}$ and $g_j^{(\mathtt{T})}$ such that:

$$\log p'_i + g_i^{(\mathtt{T})} < \log p'_j + g_j^{(\mathtt{T})} \tag{8}$$

where $p'_j := P^{do(I')|X_I=i}(X^{(\mathtt{T})} = j)$. Since gumbels sampled for Equation (7) and (8) are fixed, there must exist gumbels that satisfy both equations. The only difference is that an informative prior is imposed on these gumbels is different. Thus counterfactual stability is not violated due to the mixture prior and modified gumbel procedure. Combining the inequalities and re-arranging, we establish the contrapositive with regularization.

C Estimating counterfactual rewards with informative prior

Our proof largely follows Oberst and Sontag [43] and Buesing et al. [7] although with a different posterior on the Gumbel exogenous variables. We make the difference explicit in the following: $\mu^{(\mathtt{T})}$ be the behavior policy in the target environment and the corresponding trajectories denoted by $\tau^{\mu^{(\mathtt{T})}}$. Let $\pi^{(\mathtt{T})}$ be a candidate policy for which expected rewards are to be estimated and $\tau^{\pi^{(\mathtt{T})}}$ be the counterfactual trajectories using conditional posteriors $p(\mathbf{U}^{(\mathtt{T})|\tau})$ over exogenous variables $\mathbf{U}^{(\mathtt{T})}$. $\tau^{\pi^{(\mathtt{T})}}$ is a deterministic function of $\mathbf{U}^{(\mathtt{T})}$. The prior distributions over $\mathbf{U}$ are $p^{\pi}(\mathbf{U}^{(\mathtt{T})}) = p^{\mu}(\mathbf{U}^{(\mathtt{T})}) = p(\mathbf{U}^{(\mathtt{T})})$ (which remains the same as any informative prior coming from the source environment imposed in this framework). We drop the notation $(\mathtt{T})$ in the following as we are only concerned about the target environment hereon. Source distributions, if any, will be made explicit. Expected reward is then given by:

$$\begin{aligned}
E_{p^{\pi}}[\mathcal{R}(\tau)] &= \int_{\mathbf{u}} \mathcal{R}(\tau(\mathbf{u})) p^{\pi}(\mathbf{u}) d\mathbf{u} && (9)\\
&= \int_{\mathbf{u}} \mathcal{R}(\tau(\mathbf{u})) p^{\mu}(\mathbf{u}) d\mathbf{u} && (10)\\
&= \int_{\mathbf{u}} \mathcal{R}(\tau(\mathbf{u})) \bigg(\int_{\tau} p^{\mu}(\tau, \mathbf{u}) d\tau \bigg) d\mathbf{u} && (11)\\
&= \int_{\mathbf{u}} \mathcal{R}(\tau(\mathbf{u})) \bigg(\int_{\tau} p^{\mu}(\mathbf{u}|\tau) p^{\mu}(\tau) d\tau \bigg) d\mathbf{u} && (12)\\
&= \int_{\tau} \int_{\mathbf{u}} \mathcal{R}(\tau(\mathbf{u})) p^{\mu}(\mathbf{u}|\tau) p^{\mu}(\tau) d\mathbf{u} d\tau && (13)\\
&= E_{\tau^{\pi} \sim p^{\mu}(\tau)} \bigg[\int_{\mathbf{u}} \mathcal{R}(\tau(\mathbf{u})) p^{\mu}(\mathbf{u}|\tau) d\mathbf{u} \bigg] && (14)\\
&= E_{\tau^{\pi} \sim p^{\mu}(\tau)} \Big[E_{\mathbf{u} \sim p^{\mu}(\mathbf{u}|\tau)} [\mathcal{R}(\tau(\mathbf{u}))] \Big] && (15)
\end{aligned}$$

Where note that Equation (11) integrates over *observed* policies only. This allows to swap integrals in Equation (13). The key difference is that in Equation (15), for the subset of exogenous variables $\mathbf{g}^{(\mathtt{T})} \subseteq \mathbf{u}^{(\mathtt{T})}$, the posterior is inferred by incorporating the mixture prior that helps regularize from the source.

D KL-aggregation for CF-PI

For discrete action space, KL-aggregation for regularization over policy is equivalent to log-aggregation [24]. The proof here is provided for completeness. Consider the following aggregation setup over two discrete distributions:

$$\pi = \arg\min_{\pi} \lambda \mathrm{KL}(\pi \parallel \nu) + (1 - \lambda)\mathrm{KL}(\pi \parallel \pi^{(S)}) \tag{16}$$

This can be posed as a parametric minimization over the vector $\pi \in \Delta^{K-1}$ (where K is the dimensionality of the action space) as follows:

$$\begin{aligned} &\arg\min_{\pi} \lambda \langle \pi^T, \log \pi - \log \nu \rangle + \langle \pi^T, \log \pi - \log \pi^{\mathsf{S}} \rangle \\ &\text{s. t. } \pi \in \Delta^{K-1} \end{aligned} \tag{17}$$

Equation 17 is convex in π with a convex (simplex) constraint. Simply writing out the Lagrangian, provides:

$$\begin{aligned} &\arg\min_{\pi} \lambda \langle \pi^T, \log \pi - \log \nu \rangle + \langle \pi^T, \log \pi - \log \pi^{\mathsf{S}} \rangle + \mu(\sum_{k=1}^{K} \pi_k - 1) + \beta \pi \\ &\text{where } \beta \geq 0 \end{aligned} \tag{18}$$

Taking the gradient and setting to 0 yields:

$$(1 + \log pi) + \mu\mathbf{1} + \beta = \lambda \log \nu + (1 - \lambda) \log \pi^{\mathsf{S}} \tag{19}$$

If $1 + \mu\mathbf{1} + \beta = 0$, then $\log \pi = \lambda \log \nu + (1 - \lambda) \log \pi^{\mathsf{S}}$ and the simplex constraint is satisfied.

E Additional Experimental Details and Results

This section contains information about specific settings used to learn our policies using the various baseline approaches as well as the ablations and full CFPT procedure. We also present additional experimental findings in support of those presented in the main body of the paper.

E.1 Baseline Policy Learning Settings

As mentioned, we use the coarse sepsis simulator introduced by Oberst and Sontag [43] which can be found at `https://www.github.com/clinicalml/gumbel-max-scm`. We make one major deviation from their setting of the simulator in that we do not mask out the observations of a patient's glucose level. We also adjust the initialized proportion of diabetic patients included in the population used to define an experimental environment.

For all experiments and baselines, we fix the discount rate γ to 0.99 and the maximum number of iterations for each use of policy iteration to 1000. The number of trajectories in the source environment S was fixed to $10,000$ and the proportion of diabetic patients in S was set to 0.1. All target environments T, independent of the size of the diabetic subpopulation, were represented with 2000 trajectories. Recall that any indication of whether a patient has diabetes or not is unobserved.

In the following subsections, we report any additional parameter settings or adjustments to the learning procedure. All policy learning is done via Policy Iteration (augmented as described in the paper) utilizing an adjusted version of the `pymdptoolbox` library. Code to replicate our experiments will be made available upon publication of our paper.

E.1.1 Baselines

RANDOM This baseline doesn't explicitly learn a policy. For evaluation, all action selection is done by uniformly sampling between the 8 possible actions.

SCRATCH This non-transfer baseline constructs an empirical transition matrix from the observed data $\mathcal{H}^{(\mathsf{T})}$ which is then used within policy iteration to produce the policy $\pi^{(\mathsf{T})}$.

POOLED To pool the data between the environments S and T we estimate the transition statistics using both $\mathcal{H}^{(\mathsf{S})}$ and $\mathcal{H}^{(\mathsf{T})}$ which is then used to learn a policy with Policy Iteration in the target environment T.

BLIND This naive transfer baseline does not learn a new policy, rather it blindly uses the policy $\pi^{(\mathsf{S})}$ from the source environment without any adaptation or fine tuning. In evaluation within the target environment, actions are selected according to the distribution put forward by the source policy.

Table 1: Best performing hyperparameter settings for CFPT across each target environment T

Diabetic Proportion	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$w^{(\mathtt{T})}$	0.8	0.8	0.8	0.6	0.7	0.8	0.8	0.6	0.8	0.7	0.8
η	0.7	0.8	0.7	0.7	0.8	0.7	0.6	0.8	0.7	0.7	0.7
λ	0.9	0.9	0.3	0.1	0.3	0.6	0.3	0.1	0.3	0.4	0.9

E.1.2 Counterfactually Guided Policy Transfer (CFPT)

CFPT When applying CFPT for learning a policy in the target environment we needed to tune several hyperparameters to set-up the best policy learning environment within a data-scarce target environment T when transferring from a fixed source environment (proportion of diabetic patients: 0.1). This involved determining the best value for the number of iterations K of CF-PI, the mixture weight for regularizing the counterfactual sampling $w^{(\mathtt{T})}$, the weighting for augmenting the observed transition statistics η, and perhaps most importantly the weight for regularizing the policy learning with λ. As $w^{(T)}$, η and λ correspond to linear combinations between two quantities, we tested each of these hyperparameters between 0 and 1 in increments of 0.1, using the learned policy's true RL performance in the target environment to compare between settings. We report the optimal settings for learning within T in each target environment (diabetic proportion of population ranging from 0 to 1 in 0.1 increments) in Table 1. For all target environments the number of iterations K for CF-PI was 50.

E.1.3 Ablations

REDUCED CFPT In this ablation of CFPT, we removed the informative prior over the transition statistics within counterfactual sampling. This effectively removes this form of regularization that makes up CFPT. All other procedures and operations within CFPT were run as normal with the same parameter settings as shown in Table 1 performing best.

REGULARIZED POLICY ITERATION (REGPI) In this ablation, we removed the sampling of counterfactual trajectories completely from CFPT. We also removed any batch sampling from $\mathcal{H}^{(T)}$, using instead the full set of observed data within T. A single run of RegPI was executed, using the top performing values for λ as reported in Table 1.

E.2 Additional Results

In this section we present additional results that we did not have space to include in the main paper as well as an important additional analysis over the separate subpopulations (diabetic vs. non-diabetic) among the patients observed in the target clinical environment. In Table 2 we present the numerical values for the comparison between CFPT and all baselines and ablations shown in Figure 2b.

Approach	**Observed Reward**	**True RL Reward**	**CF-PE Reward**
Random	0.1486 ± 0.018	-0.8238 ± 0.007	0.9062 ± 0.003
Scratch	0.1486 ± 0.018	-0.7398 ± 0.007	0.9071 ± 0.003
Pooled	0.1486 ± 0.018	-0.4808 ± 0.012	0.7462 ± 0.002
Blind	0.1486 ± 0.018	-0.3937 ± 0.013	-0.7872 ± 0.006
RegPI	0.1486 ± 0.018	-0.3345 ± 0.012	0.6648 ± 0.005
Red. CFPT	0.1486 ± 0.018	-0.2122 ± 0.010	0.6819 ± 0.005
CFPT	0.1486 ± 0.018	$\mathbf{-0.1472 \pm 0.011}$	0.7333 ± 0.004

Table 2: Numerical values corresponding the policy performance results presented in Figure 2b

E.2.1 Analysis of selecting η, affecting the augmentation of $P^{(\mathtt{T})}$

In Figure 6 we demonstrate the range of policy performance under CFPT with CF-PI when varying the parameter η. Recall from Section 3.3 that η is used to weight the augmentation of the observed transition statistics in the target environment ($P^{(\mathtt{T})}$) with those estimated from the counterfactually inferred trajectories ($\hat{P}^{(\mathtt{T})}$). In this figure we demonstrate CFPT performance for policies learned in the simulated environment with a proportion of diabetic patients being 0.8, transferring from a source environment where the diabetic proportion is 0.1. The number of iterations K of CF-PI is set to 50 and we demonstrate the effect of the policy regularization parameter λ and the parameter η which is used to incorporate the inferred empirical transition matrix $\hat{P}^{(\mathtt{T})}$ into the observed target transition matrix $P^{(\mathtt{T})}$ for use in regularized policy iteration (Algorithm 4 line 26).

What we see in Figure 6 is that there is a balance when selecting η and λ for CFPT policy learning. As λ increases, meaning we are using less of the source environment, no matter the choice of η, performance more or

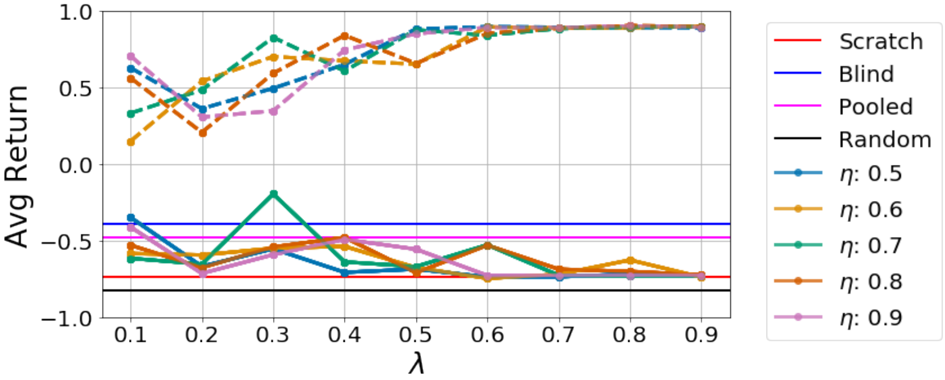

Figure 6: Demonstration of parametric study used to identify optimal settings of CFPT parameters. Shown here, within a target environment with a diabetic proportion set to 0.8 with a source population diabetic proportion set to 0.1, we see that the True RL performance (solid lines) varies as λ and η interact with a diminshed effect as λ increases. CF-PE estimated reward (dotted lines) asymptotically overestimates policy performance as λ increases.

less converges to the baseline non-transfer setting within T. However when λ is smaller, meaning we intend to use a larger proportion of the source policy, we see that the choice of η can have a broad effect. In the scenario demonstrated in Figure 6, we see that the optimal setting comes when $\eta = 0.7$ and $\lambda = 0.3$ which are the values used for all CFPT variants and ablations presented in Sec 4 when the proportion of diabetic patients in T is 0.8.

E.2.2 Sub-population Analysis of Evaluated Policies

In Figure 7 we demonstrate the differences among subpopulations when learning a policy with CFPT for different target environments T (we choose to present here the subpopulations from environments with a proportion of diabetic (pDiab) patients being 0.3, 0.5 and 0.8). When pDiab = 0.5, the performance of CFPT is only marginally better than the compared baselines. It's evaluated policy performance with CF-PE is also on par with the non-transfer baseline (SCRATCH) which is also mirrored in the aggregate counterfactual outcomes shown here as it is comparable to what has been observed when evaluating the SCRATCH baseline previously. The comparison between the two highest performing instances of CFPT (pDiab = 0.2 and pDiab = 0.8) is an interesting cross-section view of what happens when the target environment differs from the source environment. Recall that the source environment for all instances of transfer was set to pDiab = 0.1. The population of this source environment is distributionally similar to T when pDiab=0.2. Here, we see a significant increase in the number of patients who are neither discharged or die in counterfactual evaluation, in comparison to the other two pDiab settings in Figure 7. This provides some further evidence toward our conclusion that CFPT aids in the development of more circumspect policies.

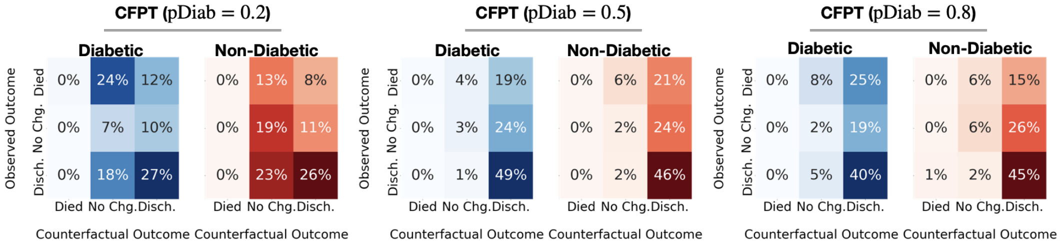

Figure 7: Aggregated counterfactual outcomes by subpopulation for different target environments T.

In Figure 8 we demonstrate the differences among subpopulations when learning a policy with CFPT having different settings of η (see Section E.2.1). With a properly chosen η (here, 0.7), we see that the evaluated outcomes of the policy increasingly push toward discharge while less optimal policies (as evaluated) appear to not have identified appropriate treatment strategies to move a majority of the observed patient trajectories toward discharge. This is most apparent when considering the non-diabetic patients, those who are in the minority within the target environment. This divergence in performance between subpopulations speaks to the importance of properly tuning the CFPT procedure.

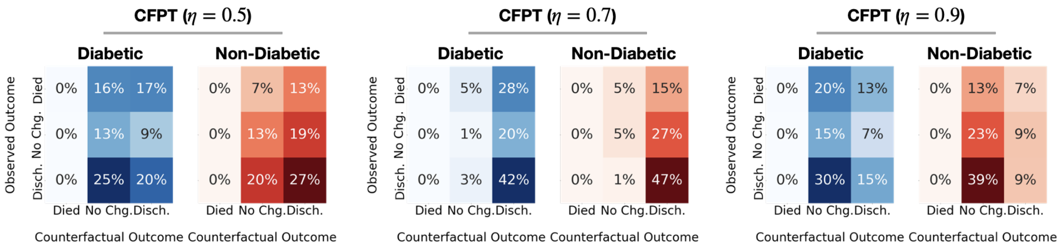

Figure 8: Aggregated counterfactual outcomes by subpopulation for different settings of η within CFPT.

Figure 9 presents an analysis between subpopulations for the non-transfer baseline (SCRATCH) and our proposed CFPT approach. Here we're looking at outcomes as inferred by counterfactual policy evaluation for the policies learned for each approach. As was discussed in Section 4.1, the policies learned via CFPT are slightly more conservative for the rarely observed non-diabetic population of the target environment. The suggested treatments and the inferred outcomes are far more measured in aggregate when using CFPT than is manifest from the non-transfer baseline.

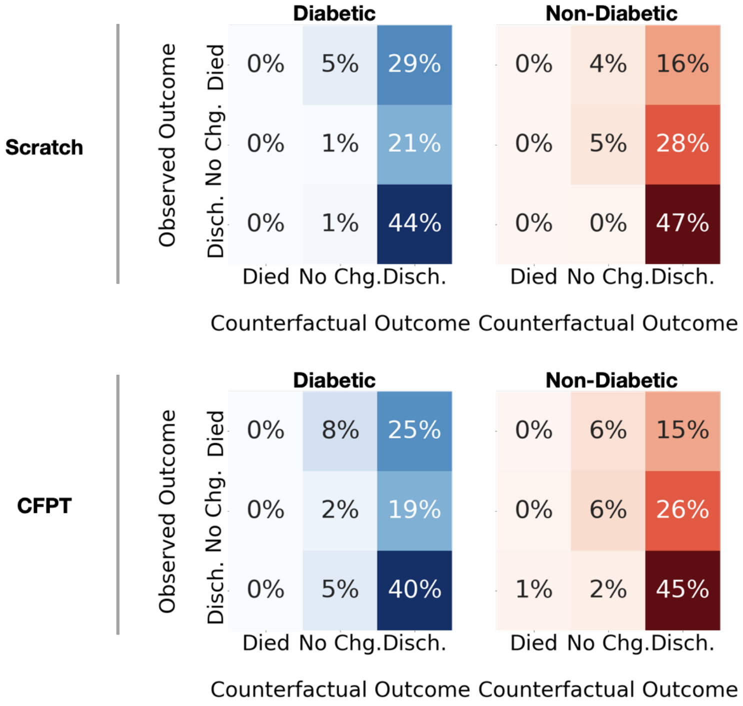

Figure 9: Aggregated counterfactual outcomes by subpopulation following the non-transfer baseline policy vs CFPT. These values are normalized by the number of patients belonging to each subpopulation (diabetic vs. non-diabetic) respectively. CFPT in aggregate is more conservative for the diabetic (rare class in source) in CF-PE evaluation.

E.2.3 Counterfactual policy evaluation: full comparison

In Figure 10 we present a full comparison between the counterfactual policy evaluation results, segmented by outcome, for each baseline and version of our proposed CFPT approach for off-policy transfer learning with limited data in the target environment. The counterfactual outcome demonstrates the unreliability of a blind transfer policy. Benefits of each parts of our regularization do shift the confidence in our policy toward discharge.

E.2.4 Introspective Analyses of Learned Policies

In this section we include additional introspective trajectory comparisons between the the non-transfer baseline (SCRATCH) and our proposed transfer procedure (CFPT). The simulated patients extracted for this comparison

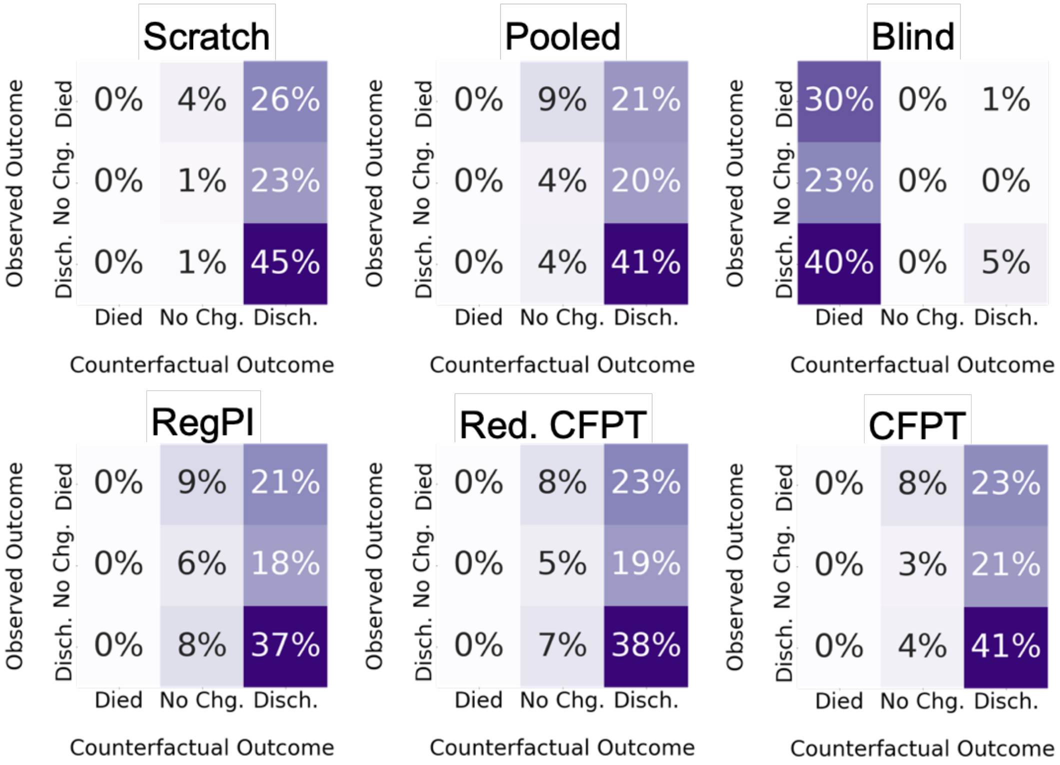

Figure 10: Comparison of all baselines in their aggregate population statistics in counterfactual evaluation of the policies learned in the target environment pDiab=0.8

are those that were observed to die where the Scratch baseline is evaluated to have treated these patients sufficiently to be discharged while CFPT is more circumspect, being evaluated to have sustained the patient's life yet not able to move them to be discharged. These examples confirm the insight reported in the main text of the paper, that the policy learned through CFPT more closely approximates the observed behavior policy in a stable fashion while also seeing slight deviations that appear to contribute to keeping the patient's vitals within a healthy range. In comparison, the non-transfer baseline policy proposes far more aggressive treatments that, in off-policy evaluation, appear to be effective yes the patient's vitals rapidly fall out of a normal or healthy range as soon as all treatments are stopped.

To augment the presentation provided in Figure 3 we include four additional trajectory introspection figures. The first of which belongs to a non-diabetic patient (Figure 11 recall, this is type of patient is found in lower proportion within the target environment) while the other three are diabetic patients (Figures 12- 14).

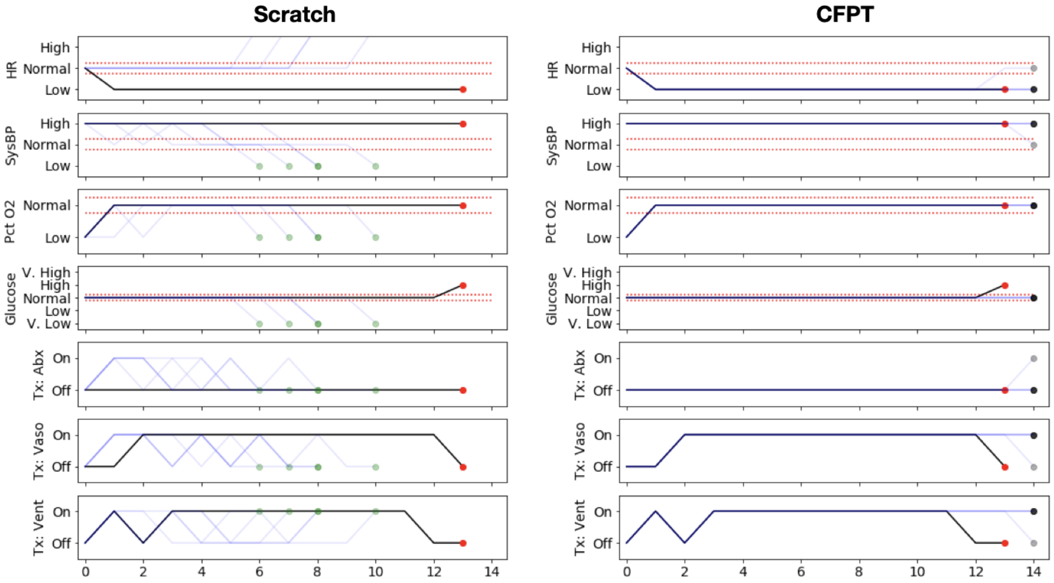

Figure 11: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is non-diabetic.

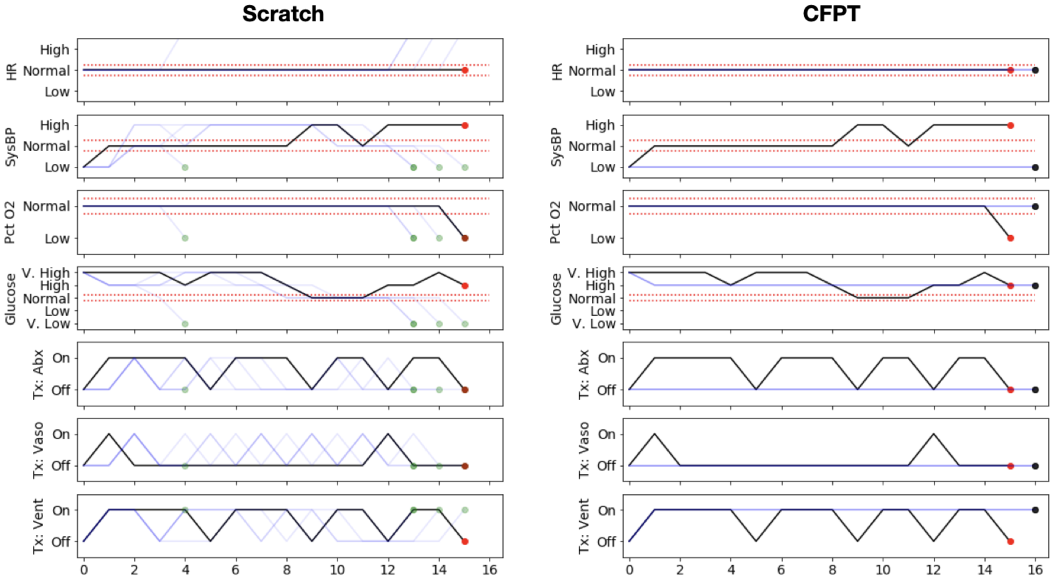

Figure 12: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is diabetic.

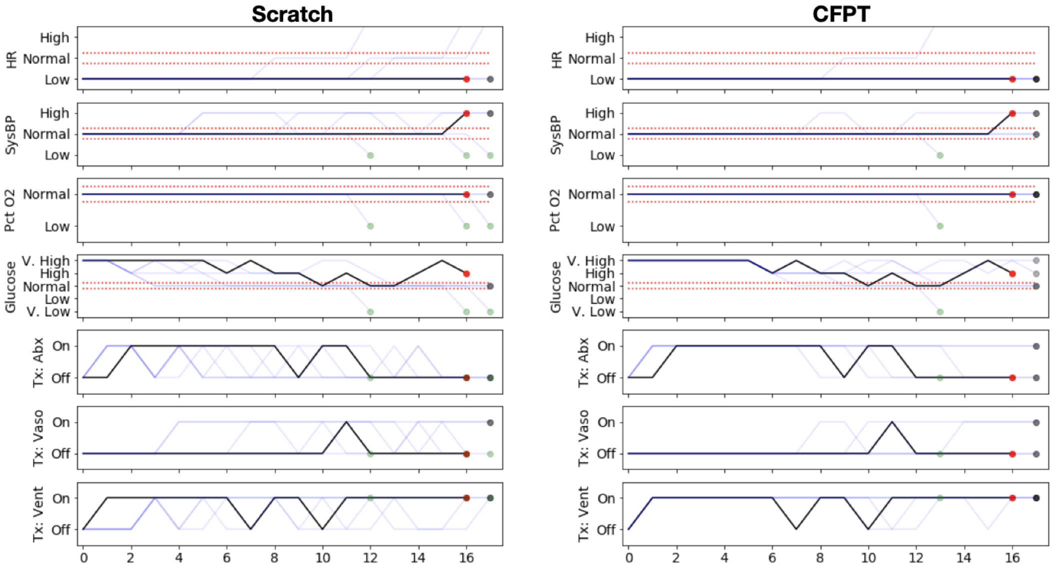

Figure 13: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is diabetic.

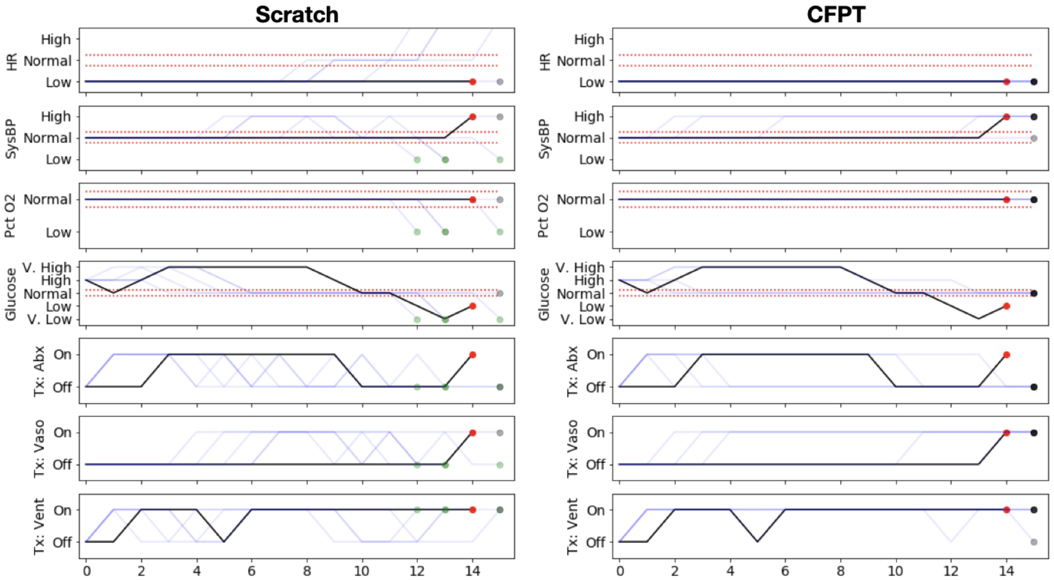

Figure 14: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is diabetic.