Counterfactually Guided Policy Transfer in Clinical Settings

Taylor W. Killian TWKILLIAN@CS.TORONTO.EDU
University of Toronto, Vector Institue, Canada

Marzyeh Ghassemi MGHASSEM@MIT.EDU
Massachusetts Institute of Technology, United States of America

Shalmali Joshi SHALMALI@SEAS.HARVARD.EDU
CRCS Harvard University (SEAS), United States of America

Abstract

Domain shift, encountered when using a trained model for a new patient population, creates significant challenges for sequential decision making in healthcare since the target domain may be both data-scarce and confounded. In this paper, we propose a method for off-policy transfer by modeling the underlying generative process with a causal mechanism. We use informative priors from the source domain to augment counterfactual trajectories in the target in a principled manner. We demonstrate how this addresses data-scarcity in the presence of unobserved confounding. The causal parametrization of our sampling procedure guarantees that counterfactual quantities can be estimated from scarce observational target data, maintaining intuitive stability properties. Policy learning in the target domain is further regularized via the source policy through KL-divergence. Through evaluation on a simulated sepsis treatment task, our counterfactual policy transfer procedure significantly improves the performance of a learned treatment policy when assumptions of "no-unobserved confounding" are relaxed.

Data and Code Availability We use data derived from a Sepsis simulator[1] to demonstrate challenges that partial observability presents when learning treatment policies (Oberst and Sontag, 2019). This simulator approximates patient physiology (discretized measurements of heart rate, blood pressure, oxygen concentration, and glucose levels) in response to medical interventions and whether the patient is diabetic. Possible treatments include antibiotics, vasopressors, and mechanical ventilation. Our code, used to augment the simulator and develop the approach described in this paper can be found at https://github.com/MLforHealth/counterfactual_transfer.

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

1. Introduction

As the development of machine learning algorithms matures there is increasing interest in deploying models to complex clinical domains (Ghassemi et al., 2018). These efforts include the application of reinforcement learning (RL) to sequential decision making and treatment recommendation (Yu et al., 2021). However, domain shift between training (source) and deployment (target) patient populations (Finlayson et al., 2021) presents challenges largely unaddressed in recent RL work. In particular, we are concerned with shifting incidence proportions of (possibly unknown and confounded) comorbidities between independent clinical environments (Subbaswamy and Saria, 2018, 2020). These challenges are amplified when few samples are available in the target domain since—for ethical and safety purposes—exploratory new data cannot be collected. Naively learned treatment policies may overfit to data-collection artefacts (Agniel et al., 2018), fail to learn meaningful interventions (François-Lavet et al., 2019), or mistime appropriate interventions (Bai et al., 2014). To provide reliable decision support and avoid such errors, principled methods are needed when transferring learned treatment policies between clinical environments.

In this paper we frame transfer in the context of offline, off-policy RL between a data-rich source domain to a data-scarce target domain, as we seek to learn robust policies from fixed observational data. We consider two main components of transfer: i) improving estimates of statistical quantities in the target domain, i.e. transition dynamics, and ii) adapting the policy learned within the source domain. We demonstrate that transfer in this setting can be naturally framed as a causal inference problem to answer the question, *"How well can a previously trained policy perform in a new target domain with limited observational data?"*

We consider the effects of data-scarcity and confounding when improving the statistical estimation of physiological responses to treatments (otherwise known as the transition dynamics) of a target patient population. Sub-populations within the observed patient cohort (perhaps categorized by disease phenotype) may exhibit dissimilar behavior in response to treatment, the composition of which may differ in target domain. When critical information about sub-populations is unavailable between domains—creating a measure of unobserved confounding and model misspecification—the accuracy of estimated dynamics will be further constrained. To address this, we propose a stochastic regularization of the estimated transition dynamics in the target domain using the estimates derived from the source domain, motivated from principles of counterfactual estimation (Pearl, 2009). We use this *counterfactual regularization* to provide a form of guided exploration in the target domain as a way to improve the estimated transition statistics.

The second component of transfer is an intelligent use of the source policy, $\pi^{(\mathsf{S})}$. Even with extensive exploration, a learned policy in the target domain may fail to converge or learn safe interventions due to regions of low data support (Gottesman et al., 2019a) or an inaccurate dynamics model (Sutton and Barto, 2018). To address this, we guide the development of the target policy $\pi^{(\mathsf{T})}$ through regularization with $\pi^{(\mathsf{S})}$. Trained with more data, $\pi^{(\mathsf{S})}$ has been exposed to a more accurate estimate of the dynamics as well as observations not present in the target domain and serves to stabilize $\pi^{(\mathsf{T})}$. By *regularizing policy learning*, we avoid undue overconfidence when determining correct treatment decisions in the target.

We propose a novel approach for policy transfer via a dual-regularization approach in offline settings. Specifically:

1. We leverage complementary elements from the source domain to support guided counterfactual sampling in the target domain which facilitates better policy learning with limited data.

2. We prove that our transfer method, Counterfactually Guided Policy Transfer (CFPT), maintains important stability properties.

3. We demonstrate, with a simulated clinical task, that CFPT obtains notable performance gains (up to 3x improvement) across domain-shifted and confounded environments.

2. Related Work

RL in Health The use of RL has been explored in healthcare to develop optimal treatment strategies (Yu et al., 2021), despite challenges presented by likely confounded data (Gottesman et al., 2019a). RL has been used to address schizophrenia (Shortreed et al., 2011), HIV (Ernst et al., 2006), sepsis (Komorowski et al., 2018; Raghu et al., 2018b; Fatemi et al., 2021) and mechanical ventilation (Prasad et al., 2017). There has also been efforts to develop reliable evaluation of learned policies since they cannot be directly tested (Kallus, 2018; Gottesman et al., 2019b; Futoma et al., 2020a) and often fail to generalize beyond their training data (Futoma et al., 2020b).

Transfer learning in RL Transfer learning in RL can improve policy learning in independent target domains (Taylor and Stone, 2009). In healthcare settings, transfer learning may enable personalized treatment strategies (Marivate et al., 2014; Killian et al., 2017) and better generalization across clinical environments. However, challenges arise as domain shift may induce additional confounding. When observations are scarce, transition estimates are prone to error (Mannor et al., 2004; Fard et al., 2008) limiting the effectiveness of counterfactual inference (the investigation of plausible alternatives to observed data). To address this, we propose a novel way to incorporate inductive bias using the source domain's transition statistics indirectly—through counterfactual inference—to leverage sub-spaces of observations that may not be in the target domain.

Causal Inference in ML Causal inference has been used to formalize counterfactual investigations of underlying data distributions (Pearl, 2009) and has recently grown to be a major focus within offline RL (Bannon et al., 2020). These foundational concepts provide benefits when addressing domain shift in supervised learning (Rojas-Carulla et al., 2018; Arjovsky et al., 2019), decision making (Makar et al., 2020; Johansson et al., 2020) and for policy reuse across multiple environments in simple bandit (Bareinboim and Pearl, 2014; Lee and Bareinboim, 2018; Lee et al., 2020) and multi-agent settings (Foerster et al., 2018). Yet, these methods require online data collection, not possible in clinical settings.

Causal concepts have also been useful evaluating policies learned from observational data (Athey, 2015; Raghu et al., 2018a) (including partially observed domains (Tennenholtz et al., 2020)). Counterfactual

reasoning in RL has been used to infer individualized treatment policies in healthcare with hidden confounding as a proxy for missing data (Parbhoo et al., 2018, 2020) or long-term effects of treatment selection (Schulam and Saria, 2017). Yet, each of these approaches rely on large and diverse training data. Our proposed transfer framework specifically relies on inducing bias (Hessel et al., 2019) indirectly by leveraging causal frameworks to incorporate an informative prior from the source domain in a partially observed sequential decision making setup.

Offline RL When learning from batch data, value function estimates to guide policy development are prone to overestimation (Hasselt, 2010) and high variance (Romoff et al., 2018). Various efforts regularize the policy learning process to maintain stability and limit extrapolation to states and actions not in the dataset (Fujimoto et al., 2019; Kumar et al., 2019). Recent offline RL algorithms additionally regularize the learned policy to remain close to observed behavior (Wu et al., 2019; Wang et al., 2020) through a KL-divergence penalty. We use a similar mechanism to constrain the target policy during learning via a form of regularized policy iteration (Farahmand et al., 2016). To the best of our knowledge, our work is the first to leverage regularized policy iteration for transfer in an offline RL setting.

3. Preliminaries

Causal modeling in RL Clinical decision making is inherently a sequential process. We model sequential decision making in this setting as a partially observed Markov decision process (POMDP) formalized by a Structural Causal Model (SCM) (Buesing et al., 2018). An SCM $\mathcal{M}$ describes the causal mechanisms of a system's observed variables $\mathbf{X}$ by defining functions $\mathbf{F}$ that govern the mechanisms, and accounting for independent stochasticity through exogenous, or external, noise variables $\mathbf{U}$. In the assumed causal graph, the nodes that directly influence a variable X_i are called the parents of X_i, $\mathbf{PA}_i$. The structural equations $f \in \mathbf{F}$ of $\mathcal{M}$ define this relationship where $X_i = f(\mathbf{PA}_i, U_i)$. Additional background is provided in the Appendix, Sec. A.

Notation To facilitate counterfactual inference for transfer from a source domain $\mathtt{S}$ to a target domain $\mathtt{T}$, we consider finite-state, finite-action episodic POMDPs. States are denoted as $S_t \in \mathcal{S}$, observations by $O_t \in \mathcal{O}$, and actions as $A_t \in \mathcal{A}$ with reward as

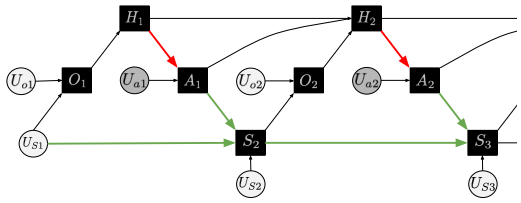

Figure 1: SCM of a POMDP from Buesing et al. (2018). White nodes denote unobserved variables, gray nodes denote observed latent variables and the black nodes are calculated quantities. We assume this structure for both the source and target domain.

$R_t = \mathcal{R}(S_t, A_t)$ for $t = \{0, 1, \ldots, T\}$. A POMDP can be represented as an SCM by expressing conditional distributions, e.g. state-transitions $P(S_{t+1}|S_t, A_t)$, as structural equations $S_{t+1} = f(S_t, A_t, U_{St})$ (Pearl, 2009), shown in Figure 1 (green edges). The relationship between A_t and the observed history $\mathcal{H}_t = \{O_1, A_1, O_2, \ldots, A_{t-1}, O_t\}$ is governed by the behavior policy μ (i.e. $\mu(A_t|H_t)$, red edges) from which trajectories $\tau = (S_1, A_1, O_1, \ldots, S_{t-1}, A_{t-1}, O_t)$ are collected with density $p^{\mu}(\tau)$.

If we choose to execute a learned policy π after having observed the behavior policy μ, the functional mapping of the red edges in Figure 1 changes from μ to π. This soft-intervention is denoted by $I'(\mu \to \pi)$ with the resulting SCM $\mathcal{M}^{do(I'(\mu\to\pi))}$. This induces a modified probability distribution, $P^{do(I'(\mu\to\pi))|I(\mu)}$ on the POMDP. We denote the corresponding "counterfactual" random variables with subscripts X_I for $do(\mu)$ and $X_{I'}$ for $do(\pi)$. The following procedure outlines how to estimate such a counterfactual distribution:

i) Abduction: estimate posteriors over exogenous noise variables $p(\mathbf{U}|\mathbf{X})$,

ii) Intervention: execute π or $do(I(\mu \to \pi))$ as if the distribution of the exogenous variables is now fixed to the posterior estimates from step i).

iii) Estimation: estimate the joint distribution of the data, will correspond with $P^{do(I'(\mu\to\pi))|I(\mu)}$.

In our case, we do not need to characterize the complete distribution of the effect a policy has on the observed data. An estimate of the reward from executing π instead of μ is sufficient. This expected reward under this counterfactual distribution can be estimated from trajectories sampled from $P^{do(I'(\mu\to\pi))|I(\mu)}$. We

denote this reward by $\mathbb{E}[\mathcal{R}(\tau)|do(\pi)]$. This expected reward can also be used to evaluate the average treatment effect under these soft-interventions to determine the value of π (see Sec. A.1.1 in the Appendix).

In general, it is not always possible to estimate $P^{do(I'(\mu\to\pi))|I(\mu)}$ and its corresponding expectations. However, we can estimate these quantities from observed data samples if we appropriately restrict the functional mappings f. One such choice of these mappings in the case of discrete or categorical states is the Gumbel-Max SCM.

Gumbel-Max Topdown Sampling. The Gumbel-Max trick enables sampling from categorical distributions $\text{Cat}(\alpha_1, \dots, \alpha_K)$, where the category k will be selected with probability α_k among K distinct categories (Hazan and Jaakkola, 2012; Maddison et al., 2014, 2016). This sampling procedure rests on inferring Gumbel variables g_k that can be transformed into these probabilities α_k.

Without any prior on the Gumbel variables $g_k^{(\mathtt{T})}$, corresponding to the discrete patient states observed in a target domain $\mathtt{T}$, the location parameters can be obtained according to the empirical transition probabilities $P^{(\mathtt{T})}$. That is, $p(\boldsymbol{\alpha}) = \delta(\log P^{\mathtt{T}}(\cdot|\cdot))$ where δ is the dirac-delta distribution. Sampling from this Gumbel given observation k' can be done using the Topdown procedure[2].

That is, for a fixed and known α_k, the Gumbel corresponding to the observed outcome k', i.e. $g_{k'}^{(\mathtt{T})}$ is itself a Gumbel variable with location parameter $Z = \log \sum_{k=1}^{K} \alpha_k$. It follows that the maximum value k' and corresponding Gumbels are independent and the rest of the exogenous variables $g_k^{(\mathtt{T})}\ \forall k \neq k'$ are truncated by this maximum value corresponding to k'. To leverage information from the source domain $\mathtt{S}$, we replace the dirac-delta prior by a mixture of the source and target transition statistics (see Sec. 4.1).

The Gumbel-Max SCM. Oberst and Sontag (2019) introduced the Gumbel-Max SCM, which ensures that counterfactual queries preserve observed outcomes (defined as *counterfactual stability*). In a Gumbel-Max SCM all nodes $\mathbf{X}$ are discrete random variables with causal mechanisms:

$$X_i \coloneqq \arg\max_j \log p(X_i = j|\mathbf{PA}_i) + g_j \tag{1}$$

given independent Gumbel variables $\mathbf{g} = \{g_1, g_2, \dots, g_k\}$. These structural equations effectively embed the Gumbel-Max trick. We parametrize the state transition mechanism of the POMDP using the formulation of Eqt. 1. This means that exogenous variables are restricted to Gumbel variables such that $U_S \triangleq \mathbf{g}$ (for all time-steps).

To ensure counterfactual expectations $\mathbb{E}[\mathcal{R}(\tau)|do(\pi)]$ are identifiable from observational data from policy μ, we need an additional property called Counterfactual Stability:

Definition 1 *Counterfactual Stability An SCM over discrete random variables is counterfactually stable if:*

$$\frac{p'_i}{p_i} \geq \frac{p'_j}{p_j} \;\Rightarrow\; P^{do(I')|X_I=i}(X = j) = 0, \quad \forall j \neq i$$

where $p_i = p(X_I = i)$ *and* $p'_i = p(X_{I'} = i)$.

Defining the structural equations in this manner acts as a constraint on the POMDP, enforcing counterfactual stability (by definition of Gumbel-Max SCMs) when considering alternative state transitions. This ensures that inferred patient outcomes change only when the relative likelihoods also change. Our counterfactual regularization further maintains this property when sampling counterfactual trajectories in the target domain after incorporating source transition estimates, outlined in Sec. 4.1. In effect all intermediate quantities remain estimable from offline data, allowing principled offline transfer.

4. Counterfactually Guided Policy Transfer

In this section we introduce a framework for transferring learned treatment policies in offline settings; meaning we only have access to sequences of observations (trajectories) τ without the ability to interact with the intended target domain. By modeling the common generative process between domains with a causal mechanism we are able to constrain policy learning in the target to refrain from unsafe behaviors, even when presented with a different and unknown mixture of patient sub-populations.

We now formalize the transfer setting. First, we assume that patients in the source and target domains have comparable health conditions. The primary shift between domains is in the composition of patient sub-populations. That is, we have different proportions of patient types (e.g. the proportion diabetic patients) in each. We assume that the population composition

2. https://cmaddis.github.io/gumbel-machinery

is unknown, creating unobserved confounding in the underlying causal system.

We assume that the data has been collected previously in the source domain $\mathtt{S}$ with some unknown behavioral policy μ and that an optimal treatment policy $\pi^{(\mathtt{S})}$ has been learned. Further, we assume that the empirical transition matrix, $P^{(\mathtt{S})}$, is accessible. Empirical transition statistics $P^{(\mathtt{T})}$ in the target are also available. We demonstrate that a learned treatment policy in $\mathtt{T}$ can be improved by 1) appropriately leveraging $P^{(\mathtt{S})}$ to improve counterfactual transition estimation in $\mathtt{T}$ and 2) regularizing $\pi^{(\mathtt{T})}$ by $\pi^{(\mathtt{S})}$.

In Sec. 4.1 we motivate that data-scarcity and unobserved confounding in $\mathtt{T}$ induces model misspecification, requiring careful regularization of the transition statistics. We propose a stochastic regularization procedure to alleviate challenges of naively transferring $P^{(\mathtt{S})}$. Our main theoretical contribution demonstrates that this procedure maintains counterfactual stability. In Sec. 4.2 we outline a second form of regularization to stabilize policy learning in $\mathtt{T}$, to avoid overconfidence in regions of little support. These concepts are combined in Sec. 4.3 to introduce our proposed transfer framework for offline, off-policy RL, Counterfactually Guided Policy Transfer (CFPT).

4.1. Counterfactual Regularization

When membership information of patient sub-populations are known, specific estimates of the transition statistics can be obtained in both domains. However, if the statistical bias in these estimates in $\mathtt{T}$ is larger for some sub-population, naive regularization from $\mathtt{S}$ can only guarantee improvement for the sub-group with more accurate estimates in $\mathtt{S}$ (see Appendix B.1).

To improve estimates of the transition statistics $P^{(\mathtt{T})}$, we need to collect more data from the appropriate counterfactual distribution i.e. $P^{do(I'(\mu\to\pi))|I(\mu)}$. Since naive regularization of $P^{(\mathtt{T})}$ is insufficient, we leverage exogenous variables in $\mathtt{T}$ (the Gumbel variables) related to $P^{(\mathtt{T})}$. According to the SCM formulation, the true posterior over these variables is completely described by the true, yet unknown, transition probabilities. Thus estimates of $P^{(\mathtt{T})}$ can be refined by improving the posterior estimates of the Gumbel variables in the "Abduction" step. The transition statistics $P^{(\mathtt{S})}$ are used to improve these posterior estimates which are then used to infer the Gumbels

Algorithm 1 Modified Top-down with informative prior

Repeat each step of a counterfactual rollout, infer τ^i
Note: - $\log P^{(\mathtt{S})}(s'|s,a) = \log \alpha^{(\mathtt{S})}$
  - $\log \hat{\alpha}^{(\mathtt{T})}$ are counterfactual stats via policy π
  - Sampled observation k'

Mixture-Topdown(SCM $\mathcal{M}$, $\log \alpha^{(\mathtt{S})}$, $\log \alpha^{(\mathtt{T})}$, $\log \hat{\alpha}^{(\mathtt{T})}$, mixture param $w^{\mathtt{T}}$, N')
// Gather a batch of counterfactual trajectories
for $n' = 1, \dots, N'$ **do**
  $\rho \sim Bernoulli(w^{(\mathtt{T})})$
  $\log \alpha = \rho \log \alpha^{\mathtt{T}} + (1-\rho) \log \alpha^{\mathtt{S}}$
  $g_{cf} = \text{Topdown}(\log \alpha, 1, k')$
  $S_{cf}^{n'} = \arg\max_j \log \hat{\alpha}^{(\mathtt{T})} + g_{cf}$
end for
$\hat{P}^{(\mathtt{T})}$ is the empirical estimate using $\{S_{cf}^{n'}\}_{n'=1}^{N'}$

in $\mathtt{T}$. This is done with a stochastic mixture of the estimated statistics from both domains.

Our key insight is that this stochastic regularization is helpful even if the mixture membership information is not known. In this case, a composite transition estimate is obtained in both $\mathtt{S}$ and $\mathtt{T}$ (instead of for each sub-population) which enables a guided sampling procedure in $\mathtt{T}$ instead of merely relying on $P^{(\mathtt{T})}$.

Concretely, we we estimate and employ the posterior $p(\mathbf{g}^{(\mathtt{S})}|\tau^{(\mathtt{S})})$ from $\mathtt{S}$ as an *informative prior* for the target domain, i.e. $p(\mathbf{g}^{(\mathtt{T})}) = p(\mathbf{g}^{(\mathtt{S})}|\tau^{(\mathtt{S})})$. This prior is incorporated in a way that maintains *counterfactual stability* in $\mathtt{T}$, allowing estimation of the expected rewards in the target domain under any candidate policy from observational data collected locally from a different policy $\mu^{\mathtt{S}}$. Normally, sampling discrete state outcomes from transition dynamics when parametrized as Gumbel-max variables leverages a sampling procedure known as Top-down sampling. This sampling procedure is a key component for estimating the expected rewards of a policy in the target domain. We ensure stability by carefully designing a modified Top-down sampling procedure (Maddison et al., 2014) when sampling from the posterior over Gumbels $\mathbf{g}^{(\mathtt{T})}$, i.e.,

$$\begin{aligned} p(\mathbf{g}^{(\mathtt{T})}|\tau^{(\mathtt{T})}, P^{(\mathtt{S})}) &\propto p(\tau^{(\mathtt{T})}|\mathbf{g}^{(\mathtt{T})}) p(\mathbf{g}^{(\mathtt{T})}) \\ &= p(\tau^{(\mathtt{T})}|\mathbf{g}^{(\mathtt{T})}) p(\mathbf{g}^{(\mathtt{S})}|P^{(\mathtt{S})}) \end{aligned}$$

given some observed trajectory $\tau^{(\mathtt{T})}$. The prior $p(\mathbf{g}^{(\mathtt{T})})$ corresponding to some state-action pair s, a is given by $p_{s,a}(\mathbf{g}^{(\mathtt{T})}) = \prod_{i=1}^{K} f_{\log P^{(\mathtt{S})}(S'=i|S,A)}(g_i)$, where $f_{\log \alpha}$ is the density of a Gumbel random variable.

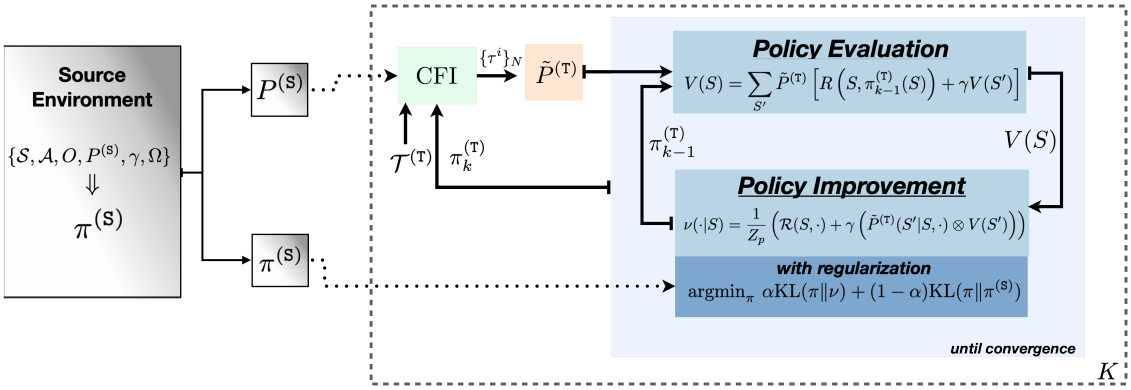

Figure 2: Graphical overview of counterfactually guided policy transfer (CFPT), as introduced in this section. Elements from the source domain are used to improve counterfactual inference (CFI) and regularize policy learning within the target domain.

To leverage the prior from $\mathtt{S}$ we impose a mixture parametrization over the *posterior* Gumbel distribution conditioned on an observation k' (in $\mathtt{T}$):

$$p(g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | k') = w^{(\mathtt{T})} p(g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | \log P^{(\mathtt{T})}, k') + w^{(\mathtt{S})} p(g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})} | \log P^{(\mathtt{S})}, k') \quad (2)$$

where $w^{(\mathtt{S})} = 1 - w^{(\mathtt{T})}$. The mixture weight w ($w < 1$) is treated as a hyper-parameter determining the amount of regularization provided by $\mathtt{S}$. This results in a modified Top-down sampling procedure, summarized in Alg. 1. Specifically, line 8 is used to select the Gumbel component from $\mathtt{S}$ or $\mathtt{T}$ with probability $w^{(\mathtt{T})}$. This component is then provided to sample the Gumbels, given observation k' from $\mathtt{T}$ (line 10). The sample is then used to infer counterfactual states under observation k', ensuring *counterfactual stability* (line 11). This modified Top-down sampling procedure provides stable counterfactual trajectories in $\mathtt{T}$ via regularization from $\mathtt{S}$ to form a batch of data to refine a treatment policy $\hat{\pi}^{(\mathtt{T})}$ from. The resulting trajectories can be used to re-estimate transition dynamics in the target domain (Alg. 2, line 7) and can also be thought of as a form of stable exploration in $\mathtt{T}$.

Lemma 2 *The mixture-prior with Modified Top-down sampling preserves counterfactual stability.*

Proof Counterfactual stability is invariant to the choice of prior so long as the gumbel samples are fixed across interventions. Our modified Top-down sampling procedure ensures this. Hence, counterfactual stability is preserved through regularization. The complete proof is in Sec. A.4 in the Appendix. ■

4.2. Regularized Policy Iteration

The sampling procedure outlined in Section 4.1 allows improved estimation of the target domain transition dynamics and evaluation of counterfactual rewards for candidate policies being considered for improvement. Policy iteration (PI) switches between evaluation and improvement steps that estimate then refine a value function V and greedy policy π. Thus, the evaluation stage of PI can leverage our modified sampling procedure. Generally, PI may not optimally converge if the MDP is partially observed (e.g. when critical sub-population information is unknown) (Sutton and Barto, 2018). When learning a policy $\pi^{(\mathtt{T})}$ in the target domain, the counterfactually sampled batch of trajectories improve the accuracy of the transition matrix used in the evaluation step of PI. However, acting greedily with respect to the inferred value function may encourage poor behavior. To guard against overconfident value estimates, we regularize the policy improvement step by $\pi^{(\mathtt{S})}$.

We regularize PI (RegPI) in $\mathtt{T}$ through minimizing the KL-divergence between the policy distributions over actions, conditioned on the observed state. Due to the discrete and finite causal framework we use to model $P^{(\mathtt{T})}$, the KL regularization is equivalent to log-aggregation (Heskes, 1998). This approach is also functionally equivalent to the behavior regularization found in recent offline RL algorithms such as BRAC (Wu et al., 2019) and CRR (Wang et al., 2020).

In this work the policies are not parametrized, so this regularization directly modifies the action distribution rather than constraining gradient updates.

Within the policy improvement step a proposal distribution $\nu(\cdot|s)$ over the actions is generated:

$$\nu(\cdot|S) = \frac{1}{Z_p}\left(\mathcal{R}(S,\cdot) + \gamma\left(\tilde{P}^{(\mathtt{T})}(S'|S,\cdot) \otimes \mathbf{V}(S')\right)\right) \quad (3)$$

where Z_p is a normalization constant and the operator $\otimes$ is used to indicate 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 the policy that minimizes the divergence between $\nu(\cdot|S)$ and $\pi^{(\mathtt{S})}(\cdot|S)$. That is,

$$\pi^{(\mathtt{T})}_{k-1} = \arg\min_{\pi} \ \lambda \ \mathrm{KL}(\pi\|\nu) + (1-\lambda)\ \mathrm{KL}(\pi\|\pi^{(\mathtt{S})}) \quad (4)$$

where λ is a hyperparameter, selected empirically to determine how much $\pi^{(\mathtt{S})}$ influences $\pi^{(\mathtt{T})}$. The derivation of Eqt. 4 and how it is fully implemented are included in the Appendix (see Sec. C and Alg. 3).

4.3. Counterfactual Policy Iteration

Algorithm 2 Counterfactual Policy Iteration

1: **CF-PI**(SCM $\mathcal{M}$, $\pi^{(\mathtt{T})}_0$, $\pi^{(\mathtt{S})}$, $P^{(\mathtt{S})}$)
2: **for** $k = 1, \ldots, K$ **do**
3: // Gather a batch of counterfactual trajectories
4: $\{h^i\}_{i=1}^N \sim \mathcal{H}^{(\mathtt{T})} \subset \mathcal{T}$
5: $\{\tau^i\}_{i=1}^N = \mathrm{CFI}(\{h^i\}_{i=1}^N, \mathcal{M}, I(\mu \to \pi^{(\mathtt{T})}_{k-1}), \mathcal{T}, P^{(\mathtt{S})})$
6: // Estimate transition stats $\hat{P}^{(\mathtt{T})}$ from $\{\tau^i\}_{i=1}^N$
7: $\tilde{P}^{(\mathtt{T})} = \frac{1}{Z_{\mathtt{T}}}\left(\eta\ P^{(\mathtt{T})} + (1-\eta)\ \hat{P}^{(\mathtt{T})}\right)$
8: // Regularized policy iteration with $\tilde{P}^{(\mathtt{T})}$
9: $\pi^{(\mathtt{T})}_k \leftarrow \mathrm{RegPI}(\pi^{(\mathtt{T})}_{k-1}, \gamma, \tilde{P}^{(\mathtt{T})}, \pi^{(\mathtt{S})}, \lambda)$
10: **end for**

We introduce counterfactually augmented policy iteration (CF-PI), the core method of our proposed CFPT framework, the major components of which have been outlined in the previous two subsections. CF-PI is visualized in Figure 2 and outlined in Alg. 2. When learning a policy in $\mathtt{T}$, where a limited number of trajectories $\mathcal{H}^{(\mathtt{T})}$ have been collected with an unknown behavior policy $\mu^{(\mathtt{T})}$, we assume access to an optimal policy distribution $\pi^{(\mathtt{S})}$ as well as transition statistics $P^{(\mathtt{S})}$ from a relevant source domain. In practice $P^{(\mathtt{S})}$ may correspond to expected patient physiological responses to treatment while $\pi^{(\mathtt{S})}$ reflects known treatment protocols.

CF-PI is performed over K iterations where, in each iteration, a batch of counterfactual trajectories $\{\tau^i\}$ from $\mathtt{T}$ (Sec. 4.1)—sampled according to the current policy $\pi^{(\mathtt{T})}_{k-1}$—are used to augment the transition statistics $P^{(\mathtt{T})}$. This augmentation (Alg. 2, line 5) is a renormalized weighted sum between the observed $P^{(\mathtt{T})}$ and $\hat{P}^{(\mathtt{T})}$ estimated from $\{\tau^i\}$. The parameter η is empirically chosen (see Sec. E.2.1 in the Appendix) to heavily favor observed transition statistics while still incorporating added diversity through counterfactual sampling. Z_T is the normalizing constant over all successor states $S' = s'$ from any given state s. $P^{(\mathtt{T})}$ is then used in regularized Policy Iteration (RegPI, Sec. 4.2) to update the policy $\pi^{(\mathtt{T})}_k$. RegPI is run to convergence or for a set number of iterations. The resulting policy $\pi^{(\mathtt{T})}_k$ is used to sample additional counterfactual trajectories at the beginning of the next iteration.

We describe the full CFPT procedure in extensive detail (including complete psuedocode) in the Appendix, Sec. D.

5. Experimental Setup

We demonstrate the benefits of CFPT through a simulated task of providing treatment to septic patients (Oberst and Sontag, 2019). We construct domain shift in the simulator by varying the proportions of diabetic patients between $\mathtt{S}$ and $\mathtt{T}$. Diabetic patients are more challenging to treat due to increased stochasticity in their glucose levels following treatment. Discharge (reward of $+1$) occurs when all vitals are 'normal' and treatment is discontinued; death (reward of -1) occurs if any three of the vitals are simultaneously not 'normal'.

Baselines:

Since generalization is guaranteed when all confounding is observed (Wen et al., 2014), we hide diabetes status, inducing unobserved confounding. This mimics realistic clinical settings where relevant information may not be immediately available. We intend to verify the robustness of CFPT in the target domain $\mathtt{T}$ even when the regularization procedure is not guaranteed to be counterfactually stable (i.e. in the presence of unobserved confounding). We compare to several baselines: i) **Scratch**, the policy $\pi^{(\mathtt{T})}$ is learned using policy iteration (PI) solely from the observed trajectories $\mathcal{H}^{(\mathtt{T})}$. ii) **Pooled** pools the observed $\mathcal{H}^{(\mathtt{S})}$ and $\mathcal{H}^{(\mathtt{T})}$ to learn $\pi^{(\mathtt{T})}$, analogous to naive regularization of $P^{(\mathtt{T})}$ by pooling data. iii) **Blind** applies $\pi^{(\mathtt{S})}$ in $\mathtt{T}$ without adaptation.

We also compare CFPT to two ablations showcasing the benefits of each contribution outlined in Sec. 4. iv) **RegPI** omits counterfactual trajectory sampling, only regularizing $\pi^{(\mathtt{T})}$ by $\pi^{(\mathtt{S})}$ (cf. Sec. 4.2), which is functionally equivalent to the tabular setting of CRR (Wang et al., 2020). v) **Red. CFPT** is a reduced form of CFPT where we omit the informative prior from $\mathtt{S}$ when sampling counterfactual trajectories. Here, the counterfactual trajectories are drawn according to the Gumbel variables from $\mathtt{T}$ only. Policy learning is then completed with RegPI. All settings used to train these policies are included in the Appendix, Sec. E.

Setup: The behavior policy μ was found using PI with full access to the MDP (including diabetes state) to provide a strong observation policy, following (Oberst and Sontag, 2019). When generating the observed trajectories $\mathcal{H}$, the policy takes random actions w.p. 0.15 to introduce variation. Within $\mathtt{S}$, $|\mathcal{H}^{(\mathtt{S})}|= 10000$, with at most 20 steps per trajectory, where the probability of a trajectory coming from a diabetic patient is 0.1. We limit $|\mathcal{H}^{(\mathtt{T})}|= 2000$ and shift the patient distribution to include a varying proportion of diabetic trajectories in range [0.0, 1.0] in 0.1 increments.

To avoid extrapolation error (Fujimoto et al., 2019) when learning $\pi^{(\mathtt{T})}$, all estimated transition statistics corresponding to actions not found in the data are zeroed out and the empirical transition matrix is renormalized. Further, the PI procedure is penalized when unsupported actions are taken; the trajectory is terminated and a negative reward is returned.

We evaluate the performance of CFPT when:

- Varying the amount of domain shift in $\mathtt{T}$, demonstrating that CFPT performs well relative to the baselines even as the patient distribution in $\mathtt{T}$ is shifted farther from $\mathtt{S}$.
- Varying the size of the patient cohort in $\mathtt{T}$, evaluating how data-scarcity affects the observed benefit of CFPT.

6. Results

6.1. CFPT is Robust Under Domain Shift

We evaluate CFPT and the baselines defined on several settings of $\mathtt{T}$ where the proportion of trajectories gathered from of diabetic patients in $\mathcal{H}^{(\mathtt{T})}$ is increased in increments of 0.1. The prevalence of diabetic patients and with few trajectories, the estimated transition statistics $P^{(\mathtt{T})}$ are far from the truth. This provides an opportunity to demonstrate the benefits of careful transfer from the source domain $\mathtt{S}$.

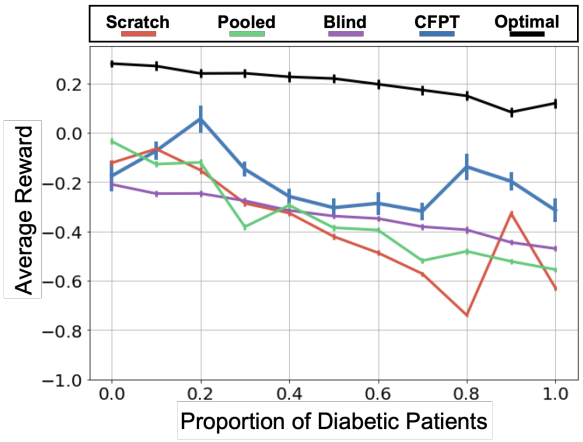

Figure 3: Comparison of CFPT with the defined baselines when varying the proportion of diabetic patients in $\mathtt{T}$. The black line denotes the observed optimal behavior with full knowledge.

6.1.1. Robustness of CFPT improvement

Figure 3 shows the average reward when applying the learned $\pi^{(\mathtt{T})}$ to simulate an additional 5000 trajectories across the various shifts in patient population in $\mathtt{T}$. The performance of $\pi^{(\mathtt{T})}$ learned with the various baseline strategies is presented alongside the observed optimal behavior $\mu^{(\mathtt{T})}$ (with full knowledge of patient state) as the solid black line. The benefits of CFPT (in blue) are clear across all levels of domain shift, with significant performance improvement when mixture populations in $\mathtt{T}$ are the furthest from $\mathtt{S}$.

Diabetic patients are harder to treat in this simulator, resulting in a decreasing trend in average reward as the proportion of diabetic patient trajectories increases. For CFPT, the advantages of leveraging $\pi^{\mathtt{S}}$ in a domain distributionally similar to $\mathtt{S}$ (pDiab= 0.3) are clear. However, in domains $\mathtt{T}$ where the patient distribution is shifted far from $\mathtt{S}$, CFPT achieves similar policy improvements. The clearest advantage of CFPT is when $\mathtt{T}$ has a majority of diabetic patient trajectories. This demonstrates that the causal framework and use of counterfactual regularization provide significant benefits when transferring from $\mathtt{S}$. Overall, this quantitative evaluation is a strong indication of the benefits of our proposed two-fold regularization when faced with domain shift between domains.

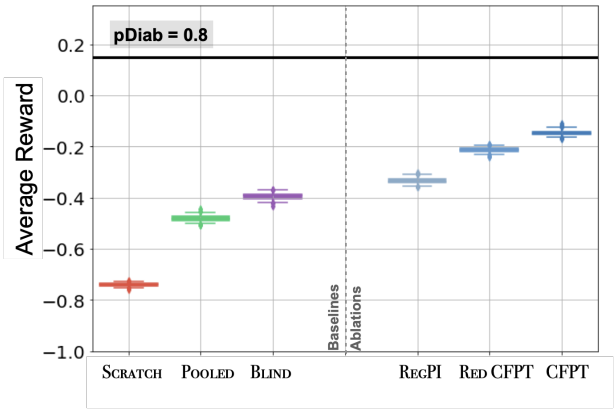

Figure 4: Comparison of estimated reward in $\mathtt{T}$ between CFPT, baselines and, ablations as outlined in Sec. 5. 95% uncertainty intervals are found via 100 bootstrapped samples of the 5000 trajectories generated with the learned policy $\pi^{(\mathtt{T})}$.

6.1.2. Ablation Study for CFPT

In Figure 4 we view the performance of CFPT, the baselines, and ablations in a setting of $\mathtt{T}$ with a 0.8 proportion of diabetic patient trajectories. The Pooled and Blind baselines provide significant improvements over Scratch. With each additional contribution we make in the development of CFPT (RegPI $\rightarrow$ Red. CFPT $\rightarrow$ CFPT) policy performance steadily improves and approaches the observed return of the optimal behavior policy derived with complete knowledge of MDP and patient diabetes state.

Recall that the RegPI ablation is functionally equivalent to the recent state of the art offline RL method CRR (Wang et al., 2020). The observed improvement over this algorithmic approach demonstrates the value of our proposed regularization for counterfactual trajectory sampling. This further validates the use of causal mechanisms when constructing a transfer approach for offline RL settings.

6.1.3. Off-policy Evaluation

The off-policy evaluation (OPE) of policies learned from fixed data, without the ability to independently test them is a challenging part of offline RL, and has been understudied in partially observed settings (Tennenholtz et al., 2020). Importance Sampling (IS) can provide an estimate of policy performance with low bias for OPE (Thomas, 2015), which is desirable in a transfer setting. While we focus on true rewards as our primary evaluation in this paper, we provide OPE estimates in this section for completeness. For this, we use weighted importance sampling (WIS) (Mahmood et al., 2014) to evaluate our transfer policies due to its consistency properties. In the Appendix, Sec. E.3.3 we use a counterfacutally determined OPE method, CF-PE (Oberst and Sontag, 2019), to qualitatively evaluate learned policies.

OPE estimates generally exhibit significant overconfidence in expected rewards, as areas of high reward are erroneously extrapolated over unseen regions of the state space. We report the results of evaluating WIS for the learned policies $\pi^{(\mathtt{T})}$ in Table 1 including comparisons to learning a policy in $\mathtt{T}$ where the diabetic status is known ("Full Obs."), through Behavior Cloning ("BC") and what the observed reward of the behavior policy $\mu^{(\mathtt{T})}$. These results are provided for the setting of $\mathtt{T}$ with a 0.8 proportion of diabetic patient trajectories. As expected, WIS overestimates the true RL return in $\mathtt{T}$, even with poor policies (i.e. **Scratch**). However, we see some semblance of improvement with each component of our proposed CFPT approach. However, the unreliability of these OPE estimates make it difficult to truly evaluate the benefits of transfer with counterfactual regularization.

Approach	True RL Reward	WIS Reward
Scratch	-0.7398 ± 0.007	0.6388 ± 0.584
Pooled	-0.4808 ± 0.012	0.9782 ± 0.004
Blind	-0.3915 ± 0.013	0.5874 ± 0.113
RegPI	-0.3366 ± 0.012	0.6266 ± 0.057
Red. CFPT	-0.2116 ± 0.010	0.7689 ± 0.077
CFPT	-0.1491 ± 0.011	0.7333 ± 0.004
Full Obs.	-0.0877 ± 0.012	0.9037 ± 0.054
BC	-0.2078 ± 0.0109	0.9836 ± 0.002
Obs. $\mu^{(\mathtt{T})}$	0.1486 ± 0.018	–

Table 1: Numerical values corresponding the policy performance results presented in Figure 4.

Fortunately, CF-PE allows for the comparison of individual counterfactual trajectories influenced by CFPT and other methods. This form of introspective evaluation can help identify glaring safety issues for deployment of a trained policy in a new domain. As seen in the Appendix, Sec. E.3.3, CFPT acts more conservatively and closely approximates the observed behavior, leading to more stable performance.

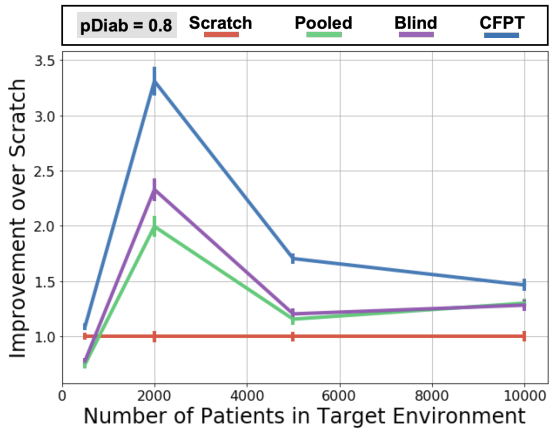

Figure 5: Performance improvement via transfer approaches over the naive Scratch baseline with respect to the number of trajectories available in T. CFPT provides significant improvement when the size of the target domain is small relative to source domain.

6.2. CFPT Demonstrates Improvement Among Various Levels of Data-Scarcity in T

The benefits of transfer may vary as more or less data is available in the observed $\mathcal{H}^{(\mathtt{T})}$. Characterizing this benefit can aid understanding of the levels of regularization one should use for transferring from S. This is particularly important when T may feature a significantly shifted data distribution, as we have simulated in this paper. Figure 5 demonstrates the improvement of different transfer approaches over a Scratch policy as the size of $\mathcal{H}^{(\mathtt{T})}$ changes. We evaluate the effects of transfer when $|\mathcal{H}^{(\mathtt{T})}| \in \{500, 2000, 5000, 10000\}$ with pDiab = 0.8. When very few samples are available, transfer does not reliably improve over Scratch, since there is little data to refine $\pi^{(\mathtt{S})}$ with. As more samples are available, clear benefits are observed from transfer, with more than a 3x improvement when using CFPT. These benefits diminish as more data is available in T, allowing for an effective policy to be learned natively. We further analyze these policy improvements for the diabetic and non-diabetic sub-populations of T in the Appendix, Sec. E.3.1.

6.3. Quality of Counterfactual Samples

We chose to use the Gumbel-Max SCM to initiate this version of our CFPT framework because it guarantees that counterfactual samples will lie within the support of $\mathcal{H}^{(\mathtt{T})}$. It is not merely a qualitative formulation, as it provides stable sampling characteristics. We quantify the quality of these counterfactual samples by comparing i) target domain samples (collected with unknown $\mu^{(T)}$) and ii) target domain counterfactual samples using $P^{(S)}$ as a prior. In Figure 6 we compare the features when diabetes status is unobserved (a corresponding analysis when the diabetes status *is* observed can be found in the Appendix, Sec. E.2.2). The counterfactually sampled data (on right) provides better coverage of the features while also not overly reducing the relative balance within the distribution of each feature. This helps to confirm the validity of using the counterfactually sampled trajectories when improving the robustness of the learned $\pi^{(\mathtt{T})}$.

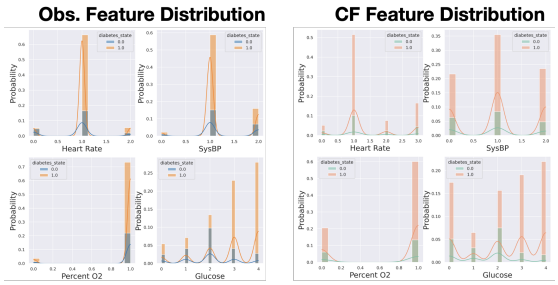

Figure 6: Feature distributions of the observed data (on left) and counterfactual samples (on right) with unobserved confounding

7. Conclusion

Motivated by challenges of policy transfer in offline, off-policy clinical settings, we have introduced Counterfactually Guided Policy Transfer. This procedure leverages complementary elements of a data-rich source domain S to facilitate better learning in a data-scarce target domain T. In our transfer framework we utilize: 1) The observed transition statistics $P^{(\mathtt{S})}$ and 2) the trained treatment policy $\pi^{(\mathtt{S})}$ to guide development of an effective policy $\pi^{(\mathtt{T})}$. By carefully designing transfer policies under restricted settings between domains we provide a principled justification for both the counterfactual and policy regularization frameworks we propose. In clinical practice, $P^{(\mathtt{S})}$ may correspond to expected patient physiological responses to treatment while $\pi^{(\mathtt{S})}$ reflects known treatment protocols. Both these elements can be feasibly shared in a secure manner and, as demonstrated by this work, used to improve treatment policy development.

In future work, we plan to adjust the regularization policies adaptively, based on the uncertainty of the transition statistics and treatment selection process. 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 domain they were 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 state spaces and treatments. This will support policy development using retrospective data derived from electronic medical records, moving us one step closer toward positively contributing to clinical practice.

Institutional Review Board (IRB)

The research presented in this paper provides a proof of concept for a novel policy transfer method and is validated on simulated data. As such this research does not require IRB approval.

Acknowledgments

We thank the anonymous reviewers and our many colleagues who contributed to thoughtful discussions and provided timely advice to improve this work. We specifically appreciate the feedback provided by Sindhu Gowda, Sana Tonekaboni, Chun-Hao Chang, Elliot Creager, David Madras, Vinith Suriyakumar and Nathan Ng.

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

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

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

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.

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.

James Bannon, Brad Windsor, Wenbo Song, and Tao Li. Causality and batch reinforcement learning: Complementary approaches to planning in unknown domains. *arXiv preprint arXiv:2006.02579*, 2020.

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.

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.

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.

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.

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

Mehdi Fatemi, Taylor W Killian, Jayakumar Subramanian, and Marzyeh Ghassemi. Medical dead-ends and learning to identify high-risk states and treatments. *Advances in Neural Information Processing Systems*, 34, 2021.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283, 2021.

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.

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.

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

Joseph Futoma, Michael C Hughes, and Finale Doshi-Velez. Popcorn: Partially observed prediction constrained reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2020a.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020b.

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

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, 2019a.

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, 2019b.

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

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

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

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

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.

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

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.

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.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11784–11794, 2019.

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

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.

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

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.

A Rupam Mahmood, Hado P van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 3014–3022, 2014.

Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In *International Conference on Machine Learning*, 2020.

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.

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.

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.

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

Sonali Parbhoo, Mario Wieser, Volker Roth, and Finale Doshi-Velez. Transfer learning from well-curated to less-resourced populations with hiv. In *Machine Learning for Healthcare Conference*, pages 589–609. PMLR, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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.

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*, 2018a.

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

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.

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.

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

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.

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.

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.

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

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

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

Philip S Thomas. Safe reinforcement learning. 2015.

Ziyu Wang, Alexander Novikov, Konrad Żołna, Jost Tobias Springenberg, Scott Reed, Bobak Shahriari, Noah Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. In *Advances in Neural Information Processing Systems*, 2020.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*. PMLR, 2014.

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

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Appendix A. Background: SCM

A.1. Structural Causal Models Pearl (2009)

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 $\mathbf{U} = \{U_1, U_2, \dots, U_k\}$ 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 $X_i \coloneqq f_i(\mathbf{PA}_i, U_i)\ \forall i$ 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 3 *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 4 *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.1.1. Connections between expected counterfactual reward and ACE/ATE

Naturally, to determine if a policy is better than a behavior policy μ, the quantity of interest is the difference in expected rewards between the behavior policy μ and another policy π. In causal inference literature, this is analogous to evaluating average treatment effect (ATE) under *soft* interventions in the underlying causal model. In our case this is a POMDP represented as a Structural Causal Model (SCM). Specifically, $\text{ATE}_\pi = E[\mathcal{R}(\tau)|do(\pi)] - E[\mathcal{R}(\tau)|do(\mu)]$ a quantity that can be interpreted as an outcome in the SCM. Note again that in off-policy settings, the first expectation term is obtained under the distribution $P^{do(I'(\mu\to\pi))|I(\mu)}$ i.e. with modified posteriors over exogenous variables.

A.2. Gumbel-Max SCM Oberst and Sontag (2019)

Definition 5 *Gumbel-Max Trick: a sampling procedure from any discrete distribution with k categories, parametrized by $p_i = P(X = i), \forall i \in \{1, 2, \dots, k\}$. 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 Pearl (2009) 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.

A.3. Gumbel-Max Topdown Sampling

The Gumbel-Max trick enables sampling from categorical distributions $\text{Cat}(\alpha_1, \ldots, \alpha_K)$, where the category k will be selected with probability α_k among K distinct categories (Hazan and Jaakkola, 2012; Maddison et al., 2014, 2016). This sampling procedure rests on inferring Gumbel variables g_k that can be transformed into these probabilities α_k.

The density of Gumbel variables with location parameter $\log \alpha_k$ and scale 1 is:

$$\begin{aligned} f_{\log \alpha_k}(g_k) &= \exp(-g_k + \log \alpha_k) \exp(-\exp(-g_k + \log \alpha_k)) \\ &= \exp(-g_k + \log \alpha_k) F_{\log \alpha_k}(g_k), \end{aligned} \tag{5}$$

where $F_{\log \alpha_k}(g_k)$ is the CDF of the Gumbel variable g_k. Without any prior on the Gumbel variables $g_k^{(\mathtt{T})}$, corresponding to the discrete patient states observed in a target domain $\mathtt{T}$, the location parameters can be obtained according to the empirical transition probabilities $P^{(\mathtt{T})}$. That is, $p(\boldsymbol{\alpha}) = \delta(\log P^{\mathtt{T}}(\cdot|\cdot))$ where δ is the dirac-delta distribution. Sampling from this Gumbel given observation k' can be done using the Topdown procedure from Maddison et al. (2014).

Now consider the joint distribution of k' and $\mathbf{g}^{(\mathtt{T})}$ for any fixed state-action pair (we drop explicit notation for clarity). To account for the informative prior $P^{(\mathtt{S})}$, we treat the locations of these Gumbel variables to be random-variables $\boldsymbol{\alpha}$. To obtain the joint distribution, we integrate over $\boldsymbol{\alpha}$:

$$p(k', g_1^{(\mathtt{T})}, \ldots, g_n^{(\mathtt{T})}) = \int_{\boldsymbol{\alpha}} \frac{\alpha_{k'}}{Z} f_{\log Z}(g_{k'}^{(\mathtt{T})}) \prod_{i \neq k'} \left[f_{\log \alpha_i}(g_i^{(\mathtt{T})}) \frac{[\![g_{k'}^{(\mathtt{T})} \geq g_i^{(\mathtt{T})}]\!]}{F_{\log \alpha_i}(g_{k'}^{(\mathtt{T})})} \right] p(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \tag{6}$$

Equation (6) can be obtained exactly following[3] Maddison et al. (2014). That is, for a fixed and known α_k, the Gumbel corresponding to the observed outcome k', i.e. $g_{k'}^{(\mathtt{T})}$ is itself a Gumbel variable with location parameter $Z = \log \sum_{k=1}^{K} \alpha_k$. It follows that the maximum value k' and corresponding Gumbels are independent and the rest of the exogenous variables $g_k^{(\mathtt{T})} \ \forall k \neq k'$ are truncated by this maximum value corresponding to k'. To leverage information from the source domain $\mathtt{S}$, we replace the dirac-delta prior by a mixture of the source and target transition statistics (see Equation 2 in Sec. 4.1). The sampling procedure follows a modified top-down procedure such that for every counterfactual sample, we first select the mixture component with probability $[w^{(\mathtt{T})}, 1 - w^{(\mathtt{T})}]$, followed by posterior sampling over the Gumbels.

A.4. Mixture-prior preserves counterfactual stability

Definition 6 *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 (2019) here:

As denoted before, let $X_I^{(\mathtt{T})} = i$ (we drop $\mathtt{T}$ from superscript for random variables when context is clear) be

3. https://cmaddis.github.io/gumbel-machinery

the outcome observed under intervention (behaviour policy) in the target domain. The state observation i implies almost surely:

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

where $p_i \coloneqq P^{(\mathbf{T})}(X = i)$ 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 $P^{do(I')|X_I^{(\mathbf{T})}=i}(X^{(\mathbf{T})} = j) \neq 0$ implies that there exist gumbel variables $g_i^{(\mathbf{T})}$ and $g_j^{(\mathbf{T})}$ such that:

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

where $p'_j \coloneqq P^{do(I')|X_I=i}(X^{(\mathbf{T})} = j)$. 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.

Appendix B. Estimating counterfactual rewards with informative prior

Our proof largely follows Oberst and Sontag (2019) and Buesing et al. (2018) although with a different posterior on the Gumbel exogenous variables. We make the difference explicit in the following: $\mu^{(\mathbf{T})}$ be the behavior policy in the target environment and the corresponding trajectories denoted by $\tau^{\mu^{(\mathbf{T})}}$. Let $\pi^{(\mathbf{T})}$ be a candidate policy for which expected rewards are to be estimated and $\tau^{\pi^{(\mathbf{T})}}$ be the counterfactual trajectories using conditional posteriors $p(\mathbf{U}^{(\mathbf{T})|\tau})$ over exogenous variables $\mathbf{U}^{(\mathbf{T})}$. $\tau^{\pi^{(\mathbf{T})}}$ is a deterministic function of $\mathbf{U}^{(\mathbf{T})}$. The prior distributions over $\mathbf{U}$ are $p^{\pi}(\mathbf{U}^{(\mathbf{T})}) = p^{\mu}(\mathbf{U}^{(\mathbf{T})}) = p(\mathbf{U}^{(\mathbf{T})})$ (which remains the same as any informative prior coming from the source environment imposed in this framework). We drop the notation $(\mathbf{T})$ 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:

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

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 $\mathbf{g}^{(\mathbf{T})} \subseteq \mathbf{u}^{(\mathbf{T})}$, the posterior is inferred by incorporating the mixture prior that helps regularize from the source.

B.1. Justification of Counterfactual Regularization

We consider two subpopulation groups (diabetic) and (non-diabetic) and the corresponding transition dynamics $P_d(S|S, A)$ and $P_{nd}(S|S, A)$. We justify our counterfactual regularization using two cases i) where diabetes status of the patient is known in both source and target environment, ii) diabetes status is unknown in both source and target. We assume here that the statistical bias in the estimated transition estimates of diabetic patients $\hat{P}_d^{(\mathbf{S})}(S|S, A)$ and $\hat{P}_d^{(\mathbf{T})}(S|S, A)$ is higher in the source domain than in the target domain (by virtue of number of samples from this subpopulation observed in both domains). The effect is the opposite for non-diabetics. i.e. the bias is lower in the source than the target domain. That is:

$$\begin{aligned}
&\|\hat{P}_d^{(\mathbf{S})}(S|S, A) - P_d(S|S, A)\| \\
&\qquad \geq \|\hat{P}_d^{(\mathbf{T})}(S|S, A) - P_d(S|S, A)\|
\end{aligned} \tag{16}$$

$$\begin{aligned}\|\hat{P}^{(\mathtt{S})}_{nd}(S|S,A) - P_{nd}(S|S,A)\| \\ \leq \ \|\hat{P}^{(\mathtt{T})}_{nd}(S|S,A) - P_{nd}(S|S,A)\|\end{aligned} \tag{17}$$

Under this setting, consider a vanilla regularization in the target-domain for the transition statistics where we use a convex combination of source and transition estimates for each sub-group instead of using the target-domain estimates only (analogously for the non-diabetic subgroup): $\eta\hat{P}^{(\mathtt{S})}_{d}(S|S,A).+(1-\eta)\hat{P}^{(\mathtt{T})}_{d}(S|S,A)$ where $0 \leq \eta \leq 1$.

Then the statistical bias for the non-diabetic group is given by:

$$\begin{aligned}&\|P_{nd}(S|S,A) - \eta\hat{P}^{(\mathtt{S})}_{nd}(S|S,A) + (1-\eta)\hat{P}^{(\mathtt{T})}_{nd}(S|S,A)\| \\ =&\| \ \eta \ (P_{nd}(S|S,A) - \hat{P}^{(\mathtt{S})}_{nd}(S|S,A)) \\ &\quad + (1-\eta)(P_{nd}(S|S,A) - \hat{P}^{(\mathtt{T})}_{nd}(S|S,A))\| \\ \leq& \ \eta \ \|\hat{P}^{(\mathtt{S})}_{nd}(S|S,A) - P_{d}(S|S,A)\| \\ &\quad + (1-\eta)\|\hat{P}^{(\mathtt{T})}_{nd}(S|S,A) - P_{nd}(S|S,A)\| \\ \leq& \ \eta \ \|\hat{P}^{(\mathtt{T})}_{nd}(S|S,A) - P_{nd}(S|S,A)\| \\ &\quad + (1-\eta)\|\hat{P}^{(\mathtt{T})}_{nd}(S|S,A) - P_{nd}(S|S,A)\| \\ =& \ \|\hat{P}^{(\mathtt{T})}_{nd}(S|S,A) - P_{nd}(S|S,A)\|\end{aligned} \tag{18}$$

The regularization from the source, done naively, will benefit the non-diabetic group. However this is not necessarily the case for the diabetic group (notice that the bias can demonstrated to be better than the source environment). However, since diabetics are the majority subpopulation in the target, such naive regularization is insufficient. Consider instead the exogenous variables corresponding to the transition dynamics model, specifically the Gumbel variables. The Gumbel variables in the source and the target are essentially parameterized by the $\log \hat{P}^{(\mathtt{S})}_{d}(S|S,A)$ and $\log \hat{P}^{(\mathtt{T})}_{d}(S|S,A)$ respectively (similarly for the non-diabetic population when the status is known). Intuitively we are essentially replacing the deterministic regularization above with a stochastic one where the so that the sampled Gumbels can still be utilized under the *true* dynamics of the target domain to generate counterfactual trajectories. Thus, our Mixture-top-down sampling can be considered as a variational/stochastic procedure to the naive regularization procedure. Notably, the stochastic procedure *decouples* the transition dynamics regularization into two steps, i) sampling Gumbels with potentially biased transition estimates, and ii) augmenting trajectories according to the true target dynamics that improves statistical estimation of the dynamics in the target.

These same insights hold true when diabetes status is not known i.e. in the presence of unobserved confounding, except that a cumulative transition statistic is available instead of separate estimates for each subpopulation.

Appendix C. KL-aggregation for CF-PI

For discrete action space, KL-aggregation for regularization over policy is equivalent to log-aggregation Heskes (1998). The proof here is provided for completeness. Consider the following aggregation setup over two discrete distributions:

$$\pi = \arg\min_{\pi} \lambda \mathrm{KL}(\pi \parallel \nu) + (1-\lambda)\mathrm{KL}(\pi \parallel \pi^{(S)}) \tag{19}$$

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:

$$\begin{aligned}&\arg\min_{\pi} \lambda\langle \pi^T, \log\pi - \log\nu\rangle + \langle \pi^T, \log\pi - \log\pi^{\mathtt{S}}\rangle \\ &\text{s. t. } \pi \in \Delta^{K-1}\end{aligned} \tag{20}$$

Equation 20 is convex in π with a convex (simplex) constraint. Simply writing out the Lagrangian, provides:

$$\begin{aligned}\arg\min_{\pi} \ \lambda \ \langle \pi^T, \log\pi - \log\nu\rangle + \langle \pi^T, \log\pi - \log\pi^{\mathtt{S}}\rangle \\ + \ \mu \ (\sum_{k=1}^{K}\pi_k - 1) + \beta\pi \\ \text{where } \beta \geq 0\end{aligned} \tag{21}$$

Taking the gradient and setting to 0 yields:

$$(1+\log\pi) + \mu\mathbf{1} + \beta = \lambda\log\nu + (1-\lambda)\log\pi^{\mathtt{S}} \tag{22}$$

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

Algorithm 3 Counterfactually Guided Policy Transfer

1: // Counterfactual inference (CFI) with source environment prior
2: **CFI**(data $\hat{x}_o$, SCM $\mathcal{M}$, intervention I, query X_q, prior $X_P^{(\mathtt{S})}$)
3: $\hat{u} \sim p(u|\hat{x}_o)$ {Sample noise variables from posterior over latent parameters}
4: $p(u) \leftarrow \delta(u - \hat{u})$ {Replace noise distribution in p with $\hat{u}$}
5: $f_i \leftarrow f_i^I$ {Perform intervention I}
6: **return** $x_q \sim p^{\mathrm{do}(I)}(x_q|\hat{u})$ {Simulate from the counterfactual posterior over model $\mathcal{M}_{\hat{x}_o}^I$, Alg. 1}

7: // Regularized Policy Iteration (RegPI)
8: **RegPI**(current policy $\pi^{(\mathtt{T})}$, discount γ, aug. statistics $\tilde{P}^{(\mathtt{T})}$, source policy $\pi^{(\mathtt{S})}$, reg. param λ)
9: Initialize $V(s)$ for all $s \in \mathcal{S}$
10: **repeat**
11: **repeat**
12: **for** each $s \in \mathcal{S}$ **do**
13: $v \leftarrow V(s)$
14: $V(s) \leftarrow \sum_{s'} \tilde{P}^{(\mathtt{T})}(s'|s, \pi^{(\mathtt{T})}(s)) \left[\mathcal{R}(s, \pi^{(\mathtt{T})}(s)) + \gamma V(s')\right]$
15: **end for**
16: **until** convergence
17: **for** each $s \in \mathcal{S}$ **do**
18: $\nu(\cdot|s) \leftarrow \frac{1}{Z_p}\left(\mathcal{R}(s, \cdot) + \gamma\left(\tilde{P}^{(\mathtt{T})}(s'|s, \cdot) \otimes \mathbf{V}(s')\right)\right)$ {Gen. a proposal dist. over actions}
19: $\pi^{(\mathtt{T})}(s) \leftarrow \arg\max_a \exp\left\{\lambda \log \nu(a|s) + (1-\lambda) \log \pi^{(\mathtt{S})}(a|s)\right\}$ {KL minimization, Eq. 22}
20: **end for**
21: **until** $\pi^{(\mathtt{T})}$ converges or after MAX˙ITERATIONS

22: // Counterfactual Policy Iteration (CF-PI)
23: **CF-PI**(SCM $\mathcal{M}$, init. policy $\pi_0^{(\mathtt{T})}$, source policy $\pi^{(\mathtt{S})}$, source statistics $P^{(\mathtt{S})}$, num. iters K, num. traj samples N, mixture param η)
24: **for** $k = 1, \ldots, K$ **do**
25: // Gather a batch of counterfactually generated trajectories in the target environment
26: $\{h^i\}_{i=1}^N \sim \mathcal{H}^{(\mathtt{T})} \subset \mathcal{T}$ {Sample batch of trajectories from observed data}
27: $\{\tau^i\}_{i=1}^N = \mathrm{CFI}(\{h^i\}_{i=1}^N, \mathcal{M}, I(\mu \to \pi_{k-1}^{(\mathtt{T})}), \mathcal{T}, P^{(\mathtt{S})})$ {Counterfactual rollouts under $\pi_{k-1}^{(\mathtt{T})}$}
28: // Estimate empirical transition statistics $\hat{P}^{(\mathtt{T})}$ from $\{\tau^i\}_{i=1}^N$
29: $\tilde{P}^{(\mathtt{T})} = \frac{1}{Z_{\mathtt{T}}}\left(\eta\ P^{(\mathtt{T})} + (1-\eta)\ \hat{P}^{(\mathtt{T})}\right)$ {Augment observed environment transition statistics}
30: // Regularized policy iteration with counterfactually augmented target env. transition statistics
31: $\pi_k^{(\mathtt{T})} \leftarrow \mathrm{RegPI}(\pi_{k-1}^{(\mathtt{T})}, \gamma, \tilde{P}^{(\mathtt{T})}, \pi^{(\mathtt{S})}, \lambda)$
32: **end for**

Appendix D. CFPT Procedure

Here we present the psuedocode (Algorithm 3) 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 4.2 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. 3, 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. 3, line 25 $\rightarrow$ CFI lines 1-6). This counterfactual sampling procedure, lever-

aging the property of counterfactual stability within Gumbel-Max SCMs, is described in Sections 4.1. 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 augmented transition matrix is then passed to RegPI as discussed in Section 4.2 (line 27 $\rightarrow$ RegPI, lines 7-21).

RegPI alternates between policy evaluation and policy improvement steps. In policy evaluation (lines 11-15) where the current policy $\pi_k^{(\mathtt{T})}$ is used to refine an estimate of the underlying value function based on the observed rewards and estimated transition statistics when applying $\pi_k^{(\mathtt{T})}$. 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 $\pi^{(\mathtt{S})}$ 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 $\pi_k^{(\mathtt{T})}$ converges or a maximum number of iterations has been performed. With this updated policy, a new batch of trajectories are sampled from $\mathcal{H}^{(\mathtt{T})}$ to draw new counterfactual samples and next iteration continues to further optimize the target policy $\pi^{(T)}$.

Appendix 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 (2019) 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 $\mathtt{S}$ was fixed to $10,000$ and the proportion of diabetic patients in $\mathtt{S}$ was set to 0.1. All target environments $\mathtt{T}$, 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 $\mathcal{H}^{(\mathtt{T})}$ which is then used within policy iteration to produce the policy $\pi^{(\mathtt{T})}$.

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

Blind This naive transfer baseline does not learn a new policy, rather it blindly uses the policy $\pi^{(\mathtt{S})}$ 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.

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

$\mathtt{T}$ 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 $w^{(\mathbf{T})}$, 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 $\mathtt{T}$ in each target environment (diabetic proportion of population ranging from 0 to 1 in 0.1 increments) in Table 2. 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 2 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 $\mathtt{T}$. A single run of RegPI was executed, using the top performing values for λ as reported in Table 2.

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 3 we present the numerical values for the comparison between CFPT and all baselines and ablations shown in Figure 4.

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

In Figure 7 we demonstrate the range of policy performance under CFPT with CF-PI when varying the parameter η. Recall from Section 4.3 that η is used to weight the augmentation of the observed transition statistics in the target environment ($P^{(\mathbf{T})}$) with those estimated from the counterfactually inferred trajectories ($\hat{P}^{(\mathbf{T})}$). 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 $\hat{P}^{(\mathbf{T})}$ into the observed target transition matrix $P^{(\mathbf{T})}$ for use in regularized policy iteration (Algorithm 3 line 26).

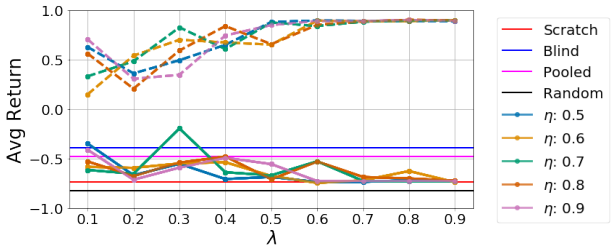

Figure 7: Demonstration of parametric study used to identify optimal settings of CFPT parameters. Shown here, within a target environment with a diabetic proportion set to 0.8 with a source population diabetic proportion set to 0.1, we see that the True RL performance (solid lines) varies as λ and η interact with a diminshed effect as λ increases. CF-PE estimated reward (dotted lines) asymptotically overestimates policy performance as λ increases.

What we see in Figure 7 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 less converges to the baseline non-transfer setting within $\mathtt{T}$. 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 7, 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 5 when the proportion of diabetic patients in $\mathtt{T}$ is 0.8.

Table 2: Best performing hyperparameter settings for CFPT across each target environment $\mathtt{T}$

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

Approach	**True RL Reward**	**WIS Reward**
Scratch	-0.7398 ± 0.007	0.6388 ± 0.584
Pooled	-0.4808 ± 0.012	0.9782 ± 0.004
Blind	-0.3915 ± 0.013	0.5874 ± 0.113
RegPI	-0.3366 ± 0.012	0.6266 ± 0.057
Red. CFPT	-0.2116 ± 0.010	0.7689 ± 0.077
CFPT	-0.1491 ± 0.011	0.7333 ± 0.004

Table 3: Numerical values corresponding the policy performance results presented in Figure 4. The observed behavior policy $\mu^{(\mathtt{T})}$ receives an average reward of 0.1486 ± 0.018.

E.2.2. Counterfactual sampling with fully observed state

Similar to the analysis presented in Section 6.3 and in Figure 6, we investigate the change in the feature distributions in $\mathtt{T}$ when the simulated patient's diabetic status is known after sampling counterfactual trajectories using the Gumbel-Max SCM, regularized by $\mathtt{S}$. The resulting comparison is shown in Figure 8.

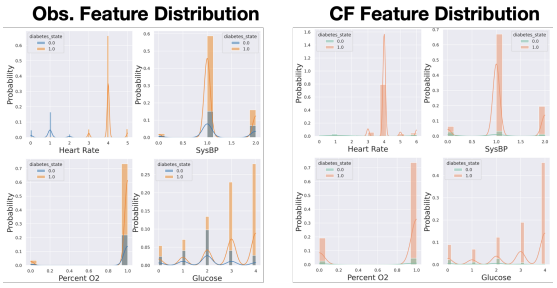

Figure 8: Feature distributions with full observations with the patient observations obtained in $\mathtt{T}$ on the left and a resampling of the feature distributions using counterfactuals drawn from the regularized Gumbel-Max SCM on the right.

E.2.3. Off-policy Evaluation of $\pi^{(\mathtt{T})}$

The off-policy evaluation (OPE) of policies learned from fixed data, without the ability to independently test them is a challenging part of offline RL, and has been understudied in partially observed settings (Tennenholtz et al., 2020). Importance Sampling (IS) can provide an estimate of policy performance with low bias for OPE (Thomas, 2015), which is desirable in a transfer setting. While we focus on true rewards as our primary evaluation in this paper, we provide OPE estimates in this section for completeness. For this, we use weighted importance sampling (WIS) (Mahmood et al., 2014) to evaluate our transfer policies due to its interesting consistency properties. In Sec. E.3.3 we use an alternative, counterfacutally determined OPE method, CF-PE (Oberst and Sontag, 2019), to qualitatively evaluate the learned policies.

OPE estimates generally exhibit significant overconfidence in expected rewards, as areas of high reward are erroneously extrapolated over unseen regions of the state space. We report the results of evaluating WIS for the learned policies $\pi^{(\mathtt{T})}$ in Table 3 for the setting of $\mathtt{T}$ with a 0.8 proportion of diabetic patient trajectories. As expected, WIS overestimates the true RL return in $\mathtt{T}$, even with poor policies (i.e. **Scratch**). However, we see some semblance of improvement with each component used to implement our proposed CFPT approach. However, the general unreliability of these OPE estimates make it difficult to truly evaluate the benefits of transfer with counterfactual regularization.

Fortunately, CF-PE allows for the comparison of individual counterfactual trajectories influenced by CFPT and other methods. This form of introspective evaluation can help identify glaring safety issues for deployment of a trained policy in a new environment. As seen in Sec. E.3.3, CFPT acts more conservatively and closely approximates the observed behavior, leading to more stable performance.

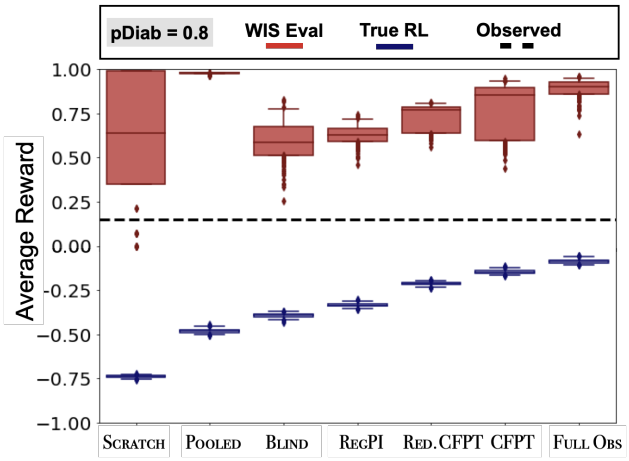

Figure 9: Comparison of estimated reward in $\mathtt{T}$ between CFPT and the baselines outlined in Sec. 5. Results after WIS are plotted in red where the true performance in $\mathtt{T}$ is plotted in blue. 95% uncertainty intervals are found through 100 bootstrapped samples of the 5000 generated trajectories under the learned target policy.

E.3. Qualitative Analysis 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 $\mathtt{T}$ (other comparisons between CFPT and the baselines are in Section E.3.3). 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 $\mathtt{T}$. In Section E.3.2 we present the aggregate counterfactual outcomes as suggested by CF-PE in comparison to what was observed. The primary difference in the evaluation between the Scratch policy and that learned through CFPT 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 10, 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.

E.3.1. Sub-population Analysis of Evaluated Policies

In Figure 11 we demonstrate the differences among subpopulations when learning a policy with CFPT for different target environments $\mathtt{T}$ (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 $\mathtt{T}$ 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 11. This provides some further evidence toward our conclusion that CFPT aids in the development of more circumspect policies.

In Figure 12 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.

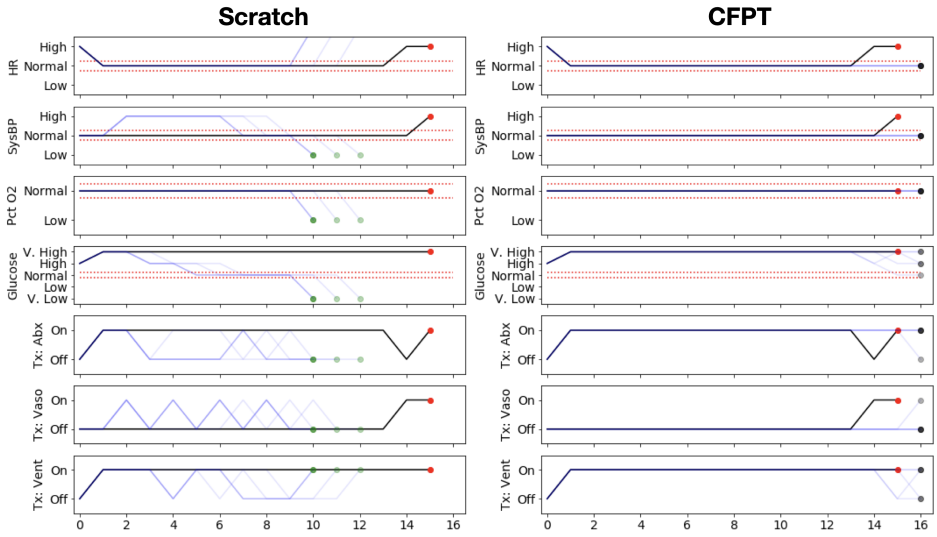

Figure 10: Qualitative comparison between CFPT and the Scratch baseline. We compare an individual patient's counterfactual trajectories using these policies. Dark lines are the observed vital measurements and actions over time while the lighter blue traces correspond to counterfactual observations and 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.3.3

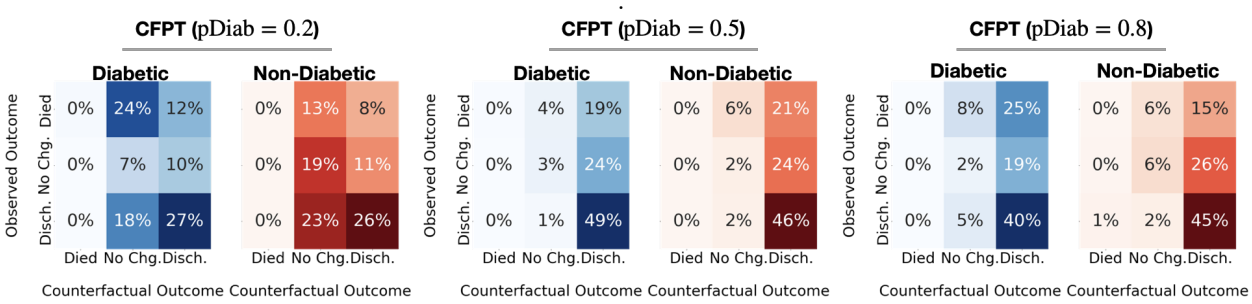

Figure 11: Aggregated counterfactual outcomes by subpopulation for different target environments τ.

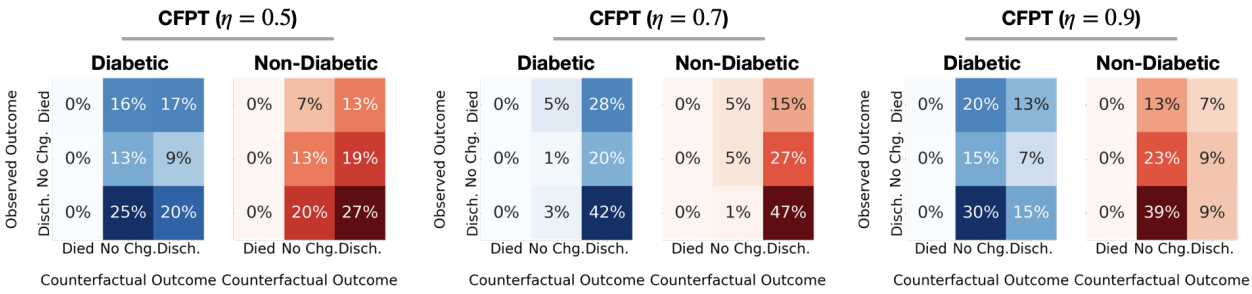

Figure 12: Aggregated counterfactual outcomes by subpopulation for different settings of η within CFPT.

Figure 13 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 6, 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.

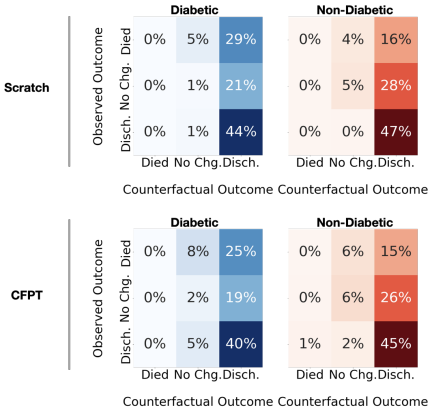

Figure 13: Aggregated counterfactual outcomes by subpopulation following the non-transfer baseline policy vs CFPT. These values are normalized by the number of patients belonging to each subpopulation (diabetic vs. non-diabetic) respectively. CFPT in aggregate is more conservative for the diabetic (rare class in source) in CF-PE evaluation.

E.3.2. Counterfactual policy evaluation: full comparison

In Figure 14 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.3.3. 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 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 10 we include four additional trajectory introspection figures. The first of which belongs to a non-diabetic patient (Figure 15 recall, this is type of patient is found in lower proportion within the target environment) while the other three are diabetic patients (Figures 16-18).

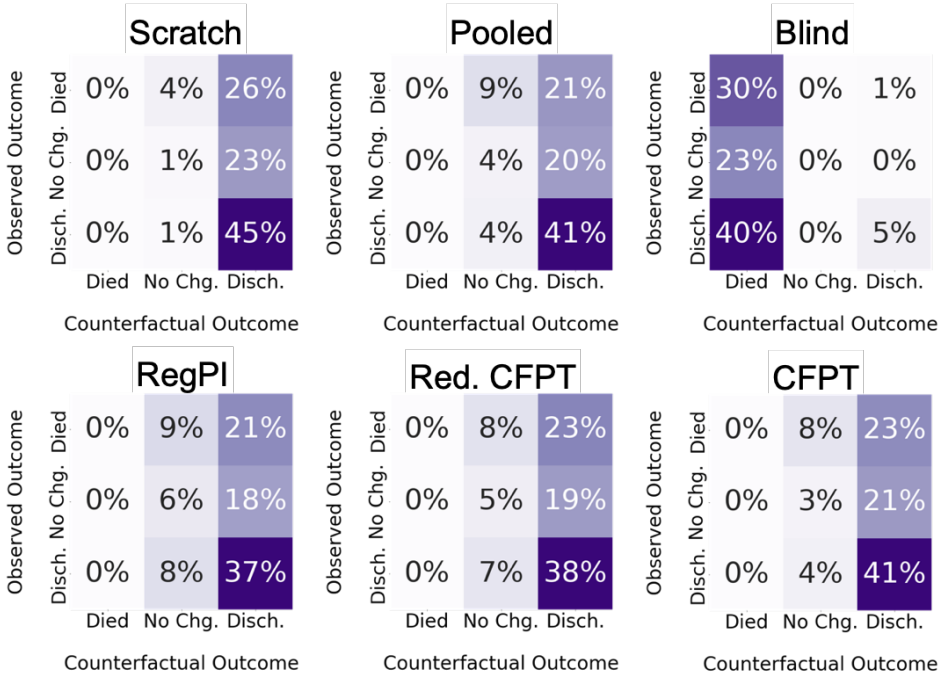

Figure 14: Comparison of all baselines in their aggregate population statistics in counterfactual evaluation of the policies learned in the target environment pDiab=0.8

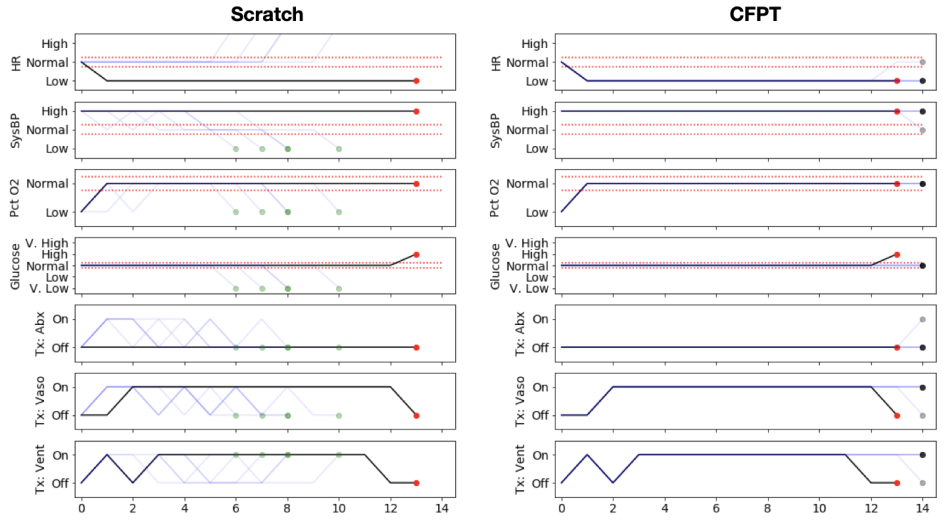

Figure 15: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is non-diabetic.

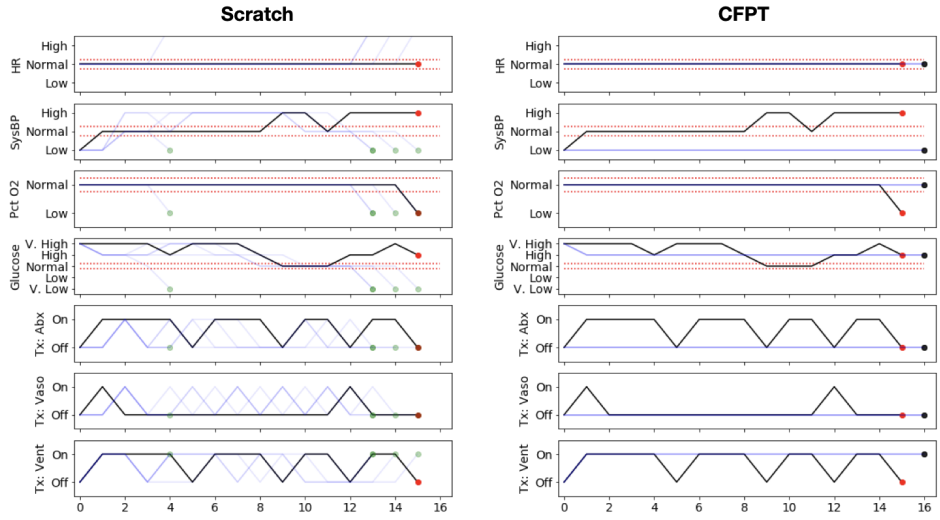

Figure 16: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is diabetic.

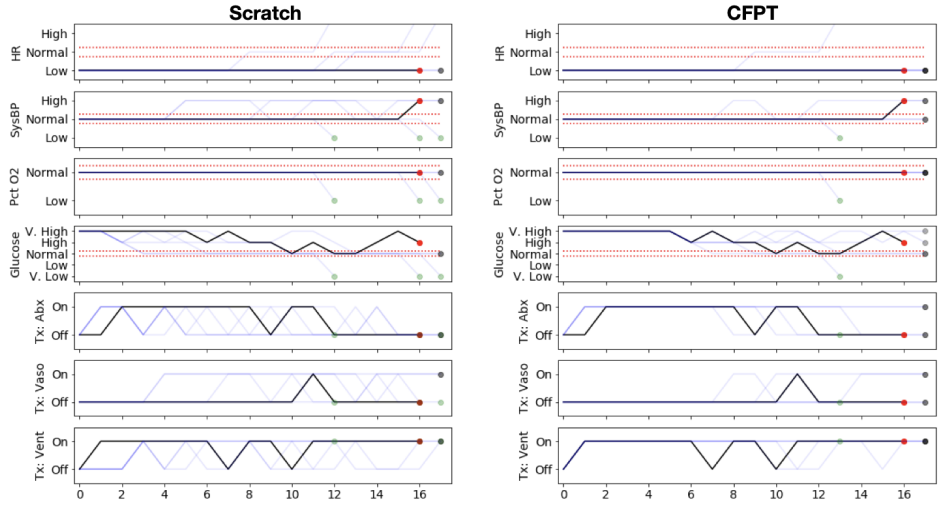

Figure 17: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is diabetic.

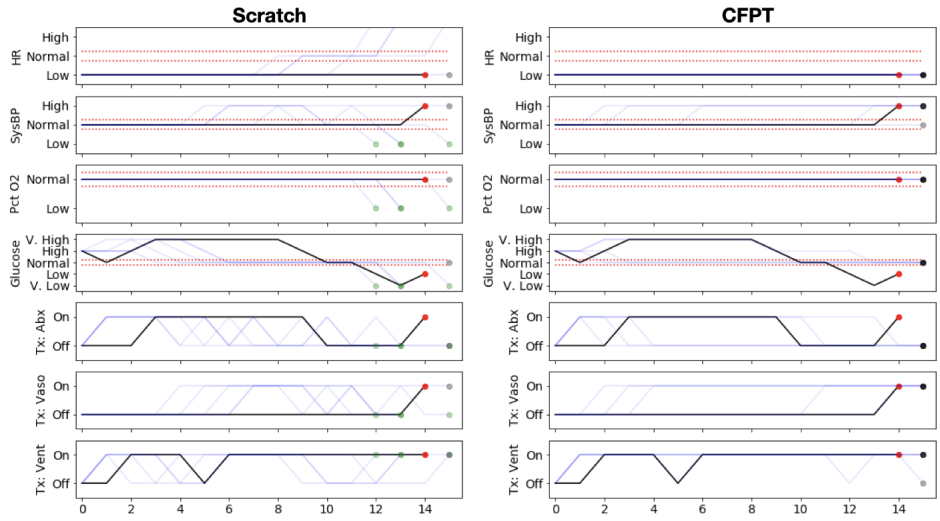

Figure 18: Introspective analysis of counterfactually sampled trajectories following the non-transfer baseline policy evaluation (left) compared with the evaluation of the proposed CFPT policy (right). This simulated patient is diabetic.