Counterfactually Guided Off-policy Transfer in Clinical Settings

Taylor W Killian∗1, Marzyeh Ghassemi†2, and Shalmali Joshi‡3

1University of Toronto, Vector Institute
2Massachussetts Institute of Technology
3CRCS, Harvard University (SEAS)

1 October, 2021

Abstract

Domain shift creates significant challenges for sequential decision making in healthcare since the target domain may be 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.

1 Introduction

With continued progress in the development of machine learning algorithms there is increasing interest in deploying models to complex, safety-critical 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., 2019]. However, domain shift between training (source) and deployment (target) domains [Panch et al., 2019, Wiens et al., 2019] presents a significant challenge largely unaddressed in recent RL work. In particular, domain shift between patient cohorts from separate clinical environments [Subbaswamy and Saria, 2018, 2020] induces confounding that needs to be addressed in order to ensure model reliability and robustness. 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. In such cases, naively learned treatment policies may significantly overfit to data-collection artefacts [Agniel et al., 2018] and fail to learn clinically meaningful interventions [François-Lavet et al., 2019]. Estimation errors due to limited data may further lead to mistimed or inappropriate interventions with adverse safety consequences [Bai et al., 2014, Waechter et al., 2014, Marik et al., 2017].

To develop generalizable models for clinical decision support, principled methods are needed to transfer treatment policies learned in data-rich source domains to data-limited target domains. In this paper, we frame transfer in the context of offline, off-policy RL, 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 would the source policy perform in the target domain with limited observational data?”

We consider the effects of data-scarcity and confounding when improving the estimation of transition statistics in the target domain. Subpopulations within the observed patient cohort (perhaps categorized by disease phenotype) may exhibit dissimilar behavior in response to treatment. When critical information about subpopulations is unavailable between domains, creating a measure of unobserved confounding and model misspecification, the accuracy of estimated transition statistics will be further constrained. In the clinical transfer setting we address in this paper, the primary variability between domains is in the subpopulation composition. That is, we have different proportions of specific patient types in each domain. Thus, the estimated dynamics between domains may differ based on the mixture distribution of subpopulations in each domain. To address this, we propose a stochastic regularization process of the estimated transition dynamics in the data-scarce target domain using the estimates derived from the data-rich source domain, motivated from principles of counterfactual estimation [Pearl, 2009]. This estimation approach mitigates harmful effects of naive alternatives to directly sharing transition statistics. 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. Note that, even with guided exploration, a policy learned in the target domain may still fail to learn safe interventions in regions of little support [Gottesman et al., 2019a] or may fail to optimally converge due to an inaccurate model of the

∗Corresponding Author: twkillian@cs.toronto.edu
†mghassem@csail.mit.edu
‡shalmali@seas.harvard.edu

1 arXiv:2006.11654v2 [cs.LG] 8 Oct 2021
domain’s dynamics [Sutton and Barto, 2018]. To counteract this, we guide the development of the target policy through regularization by the source policy. The source policy has been exposed, in the data-rich domain, to a more accurate approximation of the true dynamics as well as observations not present in the target domain and should serve to stabilize the policy in the target domain. By regularizing policy learning, we can avoid errors and undue overconfidence when determining the correct treatment decisions in the target.

We formalize these two components of regularization – to address unobserved confounding and data scarcity – in Sec. 2 to form a counterfactually guided off-policy learning algorithm in data-scarce target domains. In Sec. 3, we set up our experimental demonstration which shows notable performance gains across domain-shifted and confounded clinical environments in simulated sepsis treatment in Sec. 6. The empirical benefits of our transfer approach hold even in the presence of hidden confounding, providing up to a 3x performance improvement over directly learning a policy in the target domain. 

2 Related Work

RL in Health The use of RL has been explored in healthcare toward the development of optimal treatment strategies [Yu et al., 2019], despite challenges presented by likely confounded data [Gottesman et al., 2019a]. RL has been applied to problems addressing schizophrenia [Shortreed et al., 2011], HIV [Ernst et al., 2006], sepsis [Komorowski et al., 2018], [Raghun et al., 2018b] and mechanical ventilation [Prasad et al., 2017]. There has also been efforts to develop reliable off-policy 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 within RL can improve policy learning in independent target domains [Taylor and Stone, 2009]. Individual observations from a source domain have been previously used to directly infer rewards [Barreto et al., 2017], Laroche and Barber [2017] or accelerate policy convergence [Tirinzoni et al., 2019]. In healthcare settings, transfer learning may enable personalized treatment strategies [Marivate et al., 2014; Schulam and Saria, 2016; Kallus et al., 2017] and better generalization across clinical environments [Panch et al., 2019; Wenz et al., 2019]. However, challenges arise as domain shift may induce additional confounding. When observations are scarce, transition estimates are prone to error [Mann et al., 2004; Fard et al., 2008] limiting the effectiveness of counterfactual inference, compounding challenges of off-policy learning.

Causality 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 [Forsler 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; Raghun 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 [Parikh 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 seeks to address learning in data-limited domains by leveraging learned quantities from a data-rich source domain.

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 have sought to 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 the observed behavior policy [Wu et al., 2019; Wang et al., 2020] through a KL-divergence constraint. We use a similar mechanism to directly constrain the target policy to maintain features of the source policy during learning via a form of regularized policy iteration [Parikh et al., 2019]. 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 a natural sequential process: clinicians propose treatment, partially observe the patient response, then adapts or continues their strategy. 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 \( M \) describes the causal mechanisms of a system with an ordered triple \((U, X, F)\): independent exogenous noise variables \( U \), endogenous variables \( X \) that characterize the system, and functions \( F \) that govern the mechanisms. \( PA_i \) are the parents of \( X_i \), in the associated causal graph \( G \), i.e. the nodes that directly influence \( X_i \). The structural equations \( f \in F \) of \( M \) define this relationship where \( X_i = f(PA_i, \epsilon_i) \). Additional background is provided in the Supplement, Sec. 3.

Notation To facilitate counterfactual inference for transfer from a data-rich source to a data-scarce target, we consider finite-state, finite-action episodic POMDPs. States are denoted as \( S_t \in S \), observations by \( O_t \in O \), and actions as \( A_t \in A \) with reward as \( R_t = R(S_t, A_t) \) for \( t \in \{0,1,\ldots,T\} \). A POMDP can be represented as an
We assume that the population composition is unknown, creating unobserved confounding in the underlying

We now formalize the transfer setting. First, we assume that patients in the source and target clinical domains

This means that the exogenous variables are restricted to Gumbel variables such that

In general, it is not always possible to estimate

In our case, we do not need to characterize the complete distribution, an estimate of the reward from executing

In general, it is not always possible to estimate $P_{\text{do}(I'(\mu \rightarrow \tau))}(I(\mu))$ 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.

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 $X$ are discrete random variables with causal mechanisms:

$$X_i := \arg \max_j p(X_i = j | PA_i) + g_j$$

given independent Gumbel variables $g = \{g_1, g_2, \ldots, g_t\}$. The structural equations of the Gumbel-Max SCM effectively embed the Gumbel-Max trick [Hazan and Jaakkola (2012), Maddison et al. (2014, 2016)].

**Definition 3.1. Counterfactual Stability:** An SCM over discrete random variables is counterfactually stable if:

$$\frac{p_i'}{p_i} \geq \frac{p_j'}{p_j} \Rightarrow P_{\text{do}(I'(\mu \rightarrow \tau))}(X = j) = 0, \quad i \neq j$$

where $p_i = p(X_i = i)$ and $p'_i = p(X_i' = i)$.

where $p_i = p(X_i = i)$ and $p'_i = p(X_i' = i)$. We parametrize the state transition mechanism of the POMDP using the Gumbel-Max formulation in Eqt. 1. This means that the exogenous variables are restricted to Gumbel variables such that $U_S \triangleq g$ (for all time-steps). Defining the structural equations in this manner acts as a constraint on the POMDP, enforcing counterfactual stability when considering possible alternative state transitions. This ensures that patient outcomes change only when the relative likelihoods also change. Our counterfactual regularization maintains this property when sampling counterfactual trajectories in the target after incorporating source transition estimates as outlined in Sec. 3.1.1.

**4 Counterfactually Guided Policy Transfer**

We now formalize the transfer setting. First, we assume that patients in the source and target clinical domains have comparable health conditions. The primary variability is in the sub-population composition. That is, we have different proportions of specific patient types in each domain (e.g., the proportion diabetic patients). We assume that the population composition is unknown, creating unobserved confounding in the underlying
When membership information of sub-populations within a patient cohort is known, we can obtain specific algorithms. Algorithm 1 illustrates a modified Top-down with informative prior approach. The top-down sampling procedure [Maddison et al., 2014] is used to sample from the posterior distribution over the target POMDP. We incorporate this prior in a way that maintains counterfactual stability and, therefore, identifiability through the Gumbel-max parametrization. In Sec. 4.2, we use a second form of regularization to stabilize policy learning in the target. The policy in (1) is constrained using the source policy π(β), so as to avoid overconfidence in regions of little support. These two concepts are combined in Sec. 4.3 to outline the core of our transfer approach, Counterfactually Guided Policy Transfer (CFPT).

4.1 Counterfactual Regularization

When membership information of sub-populations within a patient cohort is known, we can obtain specific transition statistics estimates for each sub-population in both domains. However, if the statistical bias in these estimates in τ is larger for some sub-population, naive regularization from z can only guarantee improvement for the subgroup with more accurate estimates in z (see Supplement S2.1 for a justification using static regularization).

To improve estimates of the data-scarce transition statistics in τ, we need to collect more data from the appropriate counterfactual distribution i.e. \( P_{\text{data}}((\mu - \nu)|\tau)(\mu) \). Since naive regularization of transition statistics is insufficient, we leverage the exogenous variables in τ corresponding to the transition statistics i.e. the Gumbel variables. According to the SCM formulation, the true posterior over these variables is completely described by the true, yet unknown, transition probabilities \( P^{(\tau)} \). Thus estimates of \( P^{(\tau)} \) can be refined by improving the posterior estimates of the Gumbel variables in the “Abduction” step. The transition statistics \( P^{(\tau)} \) are used to improve these posterior estimates which are then used to infer the Gumbels in τ. This is done with (stochastic) mixture of the estimated statistics from both domains. As a result, the Gumbel variables in τ are sampled from a mixture of transition statistics from z and τ.

Our key insight is that this stochastic regularization is helpful even if the mixture membership information is not known (i.e. introducing unobserved confounding and thereby model mis-specification). In this case, a composite transition estimate is obtained in both z and τ (instead of for each sub-population) which enables a guided sampling procedure in τ instead of merely relying on \( P^{(\tau)} \).

Algorithm 1: Modified Top-down with informative prior

1. Repeat each step of a counterfactual rollout to infer τ^t
2. Note: - \( \log P^{(\beta)}(s'|s,a) = \log \alpha^{(\beta)}(s,a) \)
3. - \( \log \alpha^{(\tau)}(s,a) \) are counterfactual stats via the policy τ
4. - Sampled observation k^t
5. Mixture-Topdown(\( SCM, \log \alpha^{(\beta)}, \log \alpha^{(\tau)}, \log \alpha^{(\tau)}, \) mixture param \( w^T, N' \))
6. // Gather a batch of counterfactual trajectories
7. for \( n'=1,\ldots,N' \) do
8. \( \rho \sim \text{Bernoulli}(w^T) \)
9. \( \log \alpha = \log \rho \alpha^T + (1 - \rho) \log \alpha^\beta \)
10. \( g_{n,t} = \text{Topdown}(\log \alpha, 1, k^t) \)
11. \( S_{n,t}^\tau = \text{arg max}_s \log \alpha^{(\tau)}(s) + g_{n,t} \)
12. end for
13. \( P^{(\tau)} \) is the empirical estimate using \( \{S_{n,t}^\tau\}_{n,t=1}^{N' \times T} \)

Concretely, we employ the posterior \( p(g^{(\beta)}|s^{(\beta)}) \) from z as an informative prior i.e. \( p(g^{(\tau)}) = p(g^{(\beta)}|s^{(\beta)}) \) in the target POMDP. We incorporate this prior in a way that maintains counterfactual stability in τ, requiring the Gumbel samples used for intervention to be the same across all alternatives. To accomplish this, the Top-down sampling procedure [Maddison et al., 2014] is used to sample from the posterior distribution over Gumbel variables \( g^{(\tau)} \):

\[
p(g^{(\tau)}|s^{(\beta)}, P^{(\beta)}) \propto p(\tau^{(\tau)}|s^{(\beta)}, g^{(\tau)}, P^{(\beta)}) \]

\[
= p(\tau^{(\tau)}|s^{(\beta)}) p(g^{(\tau)}|P^{(\beta)})
\]

given some observed trajectory \( \tau^{(\tau)} \). The prior \( p(g^{(\tau)}|\tau^{(\tau)}) \) corresponding to some state-action pair \( s,a \) is then given by \( p_{\text{data}}(g^{(\tau)}|s,a) = \prod_{t=1}^{T} \int_{k^t \in K} \rho^{(\beta)}(s' = s'|s,a,k^t) f_{\text{Gumbel}}(k^t) \), where \( f_{\text{Gumbel}} \) is the density of a Gumbel random variable.

To leverage the informative prior from z we impose a mixture parametrization over the posterior Gumbel distribution conditioned on some observation \( k^t \) (in τ):

\[
p(g^{(\tau)}, g^{(\beta)}|k^t) \propto \frac{w^{(\tau)} p(g^{(\tau)}|s^{(\beta)}, g^{(\beta)}|k^t) p(g^{(\beta)}|k^t)}{w^{(\beta)} p(g^{(\tau)}|s^{(\beta)}, g^{(\beta)}|k^t) p(g^{(\beta)}|k^t)}
\]

where \( w^{(\beta)} = 1 - w^{(\tau)} \). The mixture weight \( w \) (\( w < 1 \)) is treated as a hyper-parameter determining the amount of regularization provided by z. This results in a modified Top-down sampling procedure, summarized...
in Alg. 1. Specifically, line 8 is used to select the Gumbel component from \( s \) or \( \tau \) with probability \( w(\tau) \). This component is then provided to Top-down sampling of the Gumbels, given observation \( k' \) from \( \tau \) (line 10). The Gumbel sample is then used to infer counterfactual states under observation \( k' \) from some policy ensuring counterfactual stability (line 11). This modified Top-down sampling procedure provides stable counterfactual trajectories in \( \tau \) via regularization from \( s \) to form a batch of data to refine a treatment policy \( \hat{\pi}(\tau) \) from. Note that while this prior in-turn induces a prior probability on the discrete distributions, the transition-estimates for policy learning are not directly regularized using this prior but via empirical estimates coming from the counterfactual trajectories within the same domain. 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 \( \tau \).

**Theorem 1.** 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. S1.4.

### 4.2 Regularized Policy Iteration

Given a fully observable, finite, and discrete MDP, Policy Iteration (PI) is guaranteed to converge to an optimal policy. PI switches between evaluation and improvement steps that estimate and refine a value function \( V \) and greedy policy \( \pi \). However, this 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 in a data-limited target domain, the counterfactually sampled batch of trajectories improve the accuracy of the transition matrix used in the evaluation step of PI (which forms \( P^T(S'|S, \cdot) \)). However, acting greedily with respect to the value function may encourage poor behavior. To guard against overconfident value estimates, we regularize the policy improvement step of PI by \( \pi^{(S)} \), which was learned from a more accurate approximation of the true dynamics.

We regularize PI (RegPI) in \( \tau \) through minimizing the KL-divergence between the policy distributions over actions, conditioned on the observed state. We are limited to discrete and finite POMDPs, due to the causal framework we use to model the state transitions. Thus the KL regularization is equivalent to log-aggregation [Heskes, 1998] and is similar 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 state-conditioned action distribution rather than constraining gradient updates.

Within the policy improvement step we generate a proposal distribution \( \nu(\cdot|s) \) over the actions:

\[
\nu(\cdot|s) = \frac{1}{Z_p} \left( R(S, \cdot) + \gamma \left( P^T(S'|S, \cdot) \circ V(S') \right) \right)
\]

(3)

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

\[
\pi_{T, k}^{(S)} = \arg \min_\pi \lambda KL(\pi(\cdot|\nu) + (1 - \lambda) KL(\pi|\pi^{(S)})
\]

(4)

where \( \lambda \) is a hyperparameter, selected empirically to determine how much \( \pi^{(S)} \) influences \( \pi^{(T)} \). The derivation of Eqn. 4 and how it is fully implemented are included in the Supplement (see Sec. S3 and Alg. 3).

### 4.3 Counterfactual Policy Iteration

We introduce counterfactually augmented policy iteration (CF-PI), the core method of our proposed CFPT procedure, 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 in \( \tau \), where a limited number of trajectories \( \pi^{(T)} \) have been collected with an unknown behavior policy \( \mu^{(T)} \), we assume access to an optimal policy distribution \( \pi^{(S)} \) as well as transition statistics \( P^{(S)} \) from a relevant source domain. In practice \( P^{(S)} \)

---

**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.
may correspond to expected patient physiological responses to treatment while $\pi^{(S)}$ reflects known treatment protocols.

CF-PI consists of $K$ iterations where, in each iteration, a batch of counterfactual trajectories ($\tau'$) from $\tau$ (Sec. 4.1) —sampled according to the current policy $\pi_{k-1}^{(T)}$— are used to augment the transition statistics $\hat{p}^{(T)}$. This augmentation (Alg. 2 line 5) is a re-normalized weighted sum between the observed $p^{(T)}$ and the estimated transition statistics $\hat{p}^{(T)}$ drawn from $\tau'$. The parameter $\eta$ is empirically chosen (see Sec. 5.2.1) to heavily favor observed transition statistics while still incorporating added diversity through counterfactual sampling. $Z_\tau$ is the normalizing constant over all successor states $S' = s'$ from any given state $s$. $\hat{p}^{(T)}$ is then used in regularized Policy Iteration (RegPI, Sec. 4.2) to update the policy $\pi_k^{(T)}$. RegPI iterates between evaluation and improvement steps, where the source policy $\pi^{(S)}$ is used to regularize the improvement step. RegPI is run to convergence or for a maximum number of iterations. The resulting policy $\pi_k^{(T)}$ 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 pseudocode, in Sec. S4.

5 Experimental Setup

We demonstrate the benefits of CFPT through a simulated task of providing treatment to septic patients. The simulator\footnote{The simulator \url{https://github.com/clinicalml/gumbel-max-scm} has been used to demonstrate challenges that partial observability presents when learning offline with limited observations.\cite{Futoma2020, Oberst2019}} approximates patient physiology (discretized measurements of heart rate, blood pressure, oxygen concentration, and glucose levels) in response to medical interventions and a latent state representing whether the patient is diabetic. Possible treatments include the binary administration of antibiotics, vasopressors, or mechanical ventilation.

We construct domain shift in the simulator by varying the proportions of diabetic patients between $\tau$ and $\tau'$. Diabetic patients are more challenging to treat due to increased stochasticity in their glucose levels following treatment. We test CFPT in a difficult transfer setting by constructing $\tau'$ to have a small number of patients while also increasing the overall prevalence of those who are diabetic. 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 \cite{Wen2014}, we hide diabetes status, inducing unobserved confounding (and thereby misspecification). This mimics realistic clinical settings where some relevant information may not be immediately available. By constructing a difficult simulated setting we intend to verify the robustness of CFPT in the shifted and data-limited $\tau$ 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 is learned using policy iteration (PI) with observational data $H^{(T)}$. ii) **Pooled** pools the observed data $H^{(S)}$ and $H^{(T)}$ to learn a policy in $\tau$. This is analogous to naive regularization of statistics by pooling data. iii) **Blind** applies $\pi^{(S)}$ in $\tau$ 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 $\pi^{(T)}$ by $\pi^{(S)}$ within PI (cf. Sec. 4.2). This serves to fine-tune $\pi^{(S)}$ in $\tau$, with the parameter $\lambda$ representing the level of adaptation, and is conceptually equivalent to the tabular setting of CBR \cite{Wang2020}. v) **Red. CFPT** is a reduced form of CFPT where we omit the informative prior from $\tau$ when sampling counterfactual trajectories. Here, the counterfactual trajectories are drawn according to the Gumbel variables from $\tau$ only. Policy learning is then completed with RegPI. All settings used to train these policies are included in Sec. S5.

**Setup:** The behavior policy $\mu$ was found using PI with full access to the MDP (including diabetes state) to provide a strong observation policy, following \cite{Oberst2019}. When generating the observed trajectories $H$, the policy takes random actions w.p. 0.15 to introduce variation. Within $\tau$, $|H^{(S)}| = 10000$, with at most 20 steps per trajectory, where the probability of a trajectory coming from a diabetic patient is 0.1. We limit $|H^{(T)}| = 2000$ to be much smaller and shift the patient distribution to include a higher proportion of diabetic trajectories. To avoid extrapolation error \cite{Fujimoto2019} when learning $\pi^{(T)}$, all estimated transition statistics corresponding to actions not found in the data are zeroed out and the empirical transition matrix is renormalized. Additionally, 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 \( T \), demonstrating that CFPT performs well relative to the baselines even as the patient distribution in \( T \) is shifted farther from \( S \).
- Varying the size of the patient cohort in \( T \), evaluating how data-scarcity affects the observed benefit of CFPT.

6 Results

6.1 CFPT is Robust Under Domain Shift

We induce domain shift in the target domain to quantitatively evaluate policy transfer using CFPT. We evaluate CFPT and the baselines defined on several settings of \( T \) where the proportion of trajectories gathered from of diabetic patients in \( H(T) \) is increased in increments of 0.1. Given the prevalence of diabetic patients and the limited number of trajectories, the estimated transition statistics \( P(T) \) will be far from the true dynamics governing the simulated care of septic patients. This provides an opportunity to demonstrate the benefits of careful transfer from the data-rich domain \( S \).

6.1.1 Robustness of CFPT improvement

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

Figure 3 shows the average reward when applying the learned \( \pi(T) \) to simulate an additional 5000 trajectories across the various shifts in patient population in \( T \). The performance of \( \pi(T) \) learned with the various baseline strategies is presented alongside the observed optimal behavior \( \mu(T) \) (with full knowledge about patient diabetes 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 \( T \) are the furthest from \( S \).

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 the source policy in a domain distributionally similar to \( S \) (\( pDiab = 0.3 \)) are clear. However, in domains \( T \) where the patient distribution is shifted far from \( S \), CFPT achieves similar policy improvements through guided perturbations via counterfactual inference. The clearest advantage of CFPT is when \( T \) has a majority of diabetic patient trajectories. This demonstrates that the causal framework and use of counterfactual regularization provide significant benefits when transferring from \( S \). Overall, this quantitative evaluation is a strong indication of the benefits of our proposed two-fold regularization when faced with domain shift between domains.

6.1.2 Ablation Study for CFPT

In Figure 4, we view the performance of CFPT, the baselines, and ablations in a setting of \( T \) with a 0.8 proportion of diabetic patient trajectories. The Pool and Blind baselines provide significant improvements over Scratch. With each additional contribution we make in the development of CFPT (RegPI → Red. CFPT → 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 conceptually 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]
We further analyze these policy improvements for the diabetic and non-diabetic sub-populations of 

The benefits of transfer may vary as more or less target data is available. Characterizing this benefit can 

Table 1:

| Approach       | True RL Reward | WIS Reward |
|----------------|----------------|------------|
| Scratch        | −0.7598 ± 0.007| 0.6888 ± 0.594 |
| 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      | −0.2136 ± 0.010| 0.7689 ± 0.077 |
| CFPT           | −0.1231 ± 0.011| 0.7334 ± 0.004 |
| Full Obs.      | −0.0877 ± 0.012| 0.9636 ± 0.004 |
| BC             | −0.2078 ± 0.0109| 0.8936 ± 0.002 |
| Obs. μ(\(T\)) | 0.1486 ± 0.018  | –          |

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

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) [Malikoud et al. 2014] to evaluate our transfer policies due to its interesting consistency properties. In Sec. S5.3 we use a counterfactually 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 \( \pi^{(T)} \) in Table 1, including comparisons to learning a policy in \( T \) where the diabetic status is known (“Full Obs.”), through Behavior Cloning (“BC”) and what the observed reward of the behavior policy \( \mu^{(T)} \). These results are provided for the setting of \( T \) with a 0.8 proportion of diabetic patient trajectories. As expected, WIS overestimates the true RL return in \( T \), 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.

Table 1: Numerical values corresponding the policy performance results presented in Figure 4. The observed behavior policy \( \mu^{(T)} \) receives an average reward of 0.1486 ± 0.019.

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 Sec. S5.3.3, CFPT acts more conservatively and closely approximates the observed behavior, leading to more stable performance.

6.2 CFPT Demonstrates Improvement Among Various Levels of Data-Scarcity in \( T \)

The benefits of transfer may vary as more or less target data is available. Characterizing this benefit can aid understanding of the levels of regularization one should use for transferring from \( T \). 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 \( T \) changes. We evaluate the effects of transfer when \(|H^{(T)}| \in \{ 500, 2000, 5000, 10000 \} \) with \( p_{\text{Diab}} = 0.8 \). When very few samples are available, transfer does not reliably improve over Scratch, since there is little data to refine \( \pi^{(T)} \) 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 an increase in the amount of data available in \( T \) allows for an effective policy to be learned natively, yet some improvement is still seen through CFPT. We further analyze these policy improvements for the diabetic and non-diabetic sub-populations of \( T \) in Sec. S5.3.1.
Figure 5: The improvement transfer approaches have over the naive Scratch baseline with respect to the number of trajectories available in $T$. CFPT provides significant improvement even as this effect wanes as the size of the target domain increases.

6.3 Quality of Counterfactual Samples

We chose to use the Gumbel-Max SCM formulation because it guarantees that counterfactual samples will lie within the support of the batch data found in the target domain. It is not merely a qualitative formulation as it provides stable sampling characteristics, ensuring that the counterfactuals are supported by the data. 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 $\pi^{(S)}$ as a prior within the Gumbel-Max SCM. In Figure 6 we compare the features when diabetes status is unobserved (a corresponding analysis when the diabetes status is observed is included in Section S5.2.2 in the Appendix). 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 regularized Gumbel-Max SCM to sample trajectories when improving the robustness of the learned policy $T$.

Figure 6: Feature distributions 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 $P^{(S)}$ and 2) the trained treatment policy $\pi^{(S)}$ to guide development of an effective policy $\pi^{(T)}$. By carefully designing transfer policies under restricted settings—like shifts to mixture subpopulations and unobserved confounding between domains—we provide a principled justification for both the counterfactual and policy regularization frameworks we propose. In clinical practice, $P^{(S)}$ may correspond to the expected patient physiological response to a prescribed treatment while $\pi^{(S)}$ reflects known treatment protocols when presented with a particular patient context. Both these elements can be feasibly shared in a secure and private manner and, as demonstrated by this work, used to improve treatment policy development.
In 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.

Acknowledgements

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

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

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

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 Barreinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Advances in neural information processing systems, pages 280–288, 2014.

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

Laes Buesing, Theophane Weber, Yuri 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 AAI, pages 1056–1061, 2008.

Jakob N Foerster, Gregory Farquhar, Tritantafilos 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:1806.09388, 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.
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-Laret, 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. Integrative analysis using coupled latent variable models for individualizing prognoses. *The Journal of Machine Learning Research*, 17(1):8244–8278, 2016.

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.

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.

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

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

Ziyu Wang, Alexander Novikov, Konrad Żołna, Josh 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.

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

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

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: a survey. *arXiv preprint arXiv:1908.08796*, 2019.
S1 Background: SCM

S1.1 Structural Causal Models [Pearl 2009]
A structural causal model $M$ describes the causal mechanisms driving a system. It consists of an ordered triple $(U, X, F)$; a set of independent exogenous random variables $U = \{U_1, U_2, \ldots, U_k\}$ that represent factors of variation outside the model, $X$ comprises the endogenous variables modeled in the causal system and, the set of functions $F$ defined by $X_i : = f_i(PA_i, U_i)$ where $PA_i \subseteq X \setminus X_i$ govern the causal mechanisms. $PA_i$ are the parents of $X_i$ in a causal DAG $G$. The framework attributes probabilistic Markov assumptions to the joint distribution $P^M$ associated with the variables $(X, U)$ in the graph. This characterizes a probability distribution, implying that one can observe samples true to the underlying causal graph and mechanism.

Definition S1.1. Interventional Distribution: An intervention $I$ in an SCM $M$ consists of replacing some functions $f_i(PA_i, U_i)$ with a different governing causal mechanism $f'_i(PA'_i, U'_i)$ where $PA'_i$ are the parents of $X_i$ in a new DAG $G'$. Note that the interventional distribution does not change the exogenous mechanisms driving the system. The resulting SCM, denoted by $M^{I^{do}(i)}$ has a new joint distribution denoted by $P^{M^{I^{do}(i)}}$.

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 $P(U | X = x)$ and intervene with $I$ on a causal system with exogenous noise priors $p(U)$ replaced by $P(U | X = x)$.

Definition S1.2. Counterfactual Distribution: Let $M^*$ correspond to the SCM where the exogenous noise model $p(U)$ in $M$ is replaced by $P(U | X = x)$. Intervening with $I$ on the resulting SCM $M^*$ yields a new SCM $M^{I^{do}(i)|I^{do}(j)}$ and induces the joint counterfactual distribution $P^{M^{I^{do}(i)|I^{do}(j)}}$.

S1.1.1 Connections between expected counterfactual reward and ACE/ATE

Naturally, to determine if a policy is better than a behavior policy $\pi$, the quantity of interest is the difference in expected rewards between the behavior policy $\pi$ and another policy $\mu$. 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, $ATE_\pi = E(R(\pi) - R(\mu))$ 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 $P^{do(I^{do}(\mu \rightarrow x))}(I^{do}(\mu))$ i.e. with modified posteriors over exogenous variables.

S1.2 Gumbel-Max SCM [Oberst and Sontag 2019]

Definition S1.3. Gumbel-Max Trick: a sampling procedure from any discrete distribution with $k$ categories, parameterized by $p_i = P(X = i), \forall i \in \{1, 2, \ldots, k\}$. First, sample $k$ independent Gumbel variables $g_j$ with location 0, scale 1. Set the sampled outcome $k = \text{arg max}_i \log p_j + g_j$.

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

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 $P^{do(I^{do}(X = i))}(Y = i) = p_i, \forall i \in [L]$ and $P^{do(I^{do}(X = i))}(Y = i) = p'_i, \forall i \in [K]$. Let $P^{do(I^{do}(X = i))}(Y = i) = p_i$ 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 $P^{do(I^{do}(X = i))}(Y = j)$ is the counterfactual probability of observing outcome $j$ having observed $i$ under intervention $I$.

S1.3 Gumbel-Max Topdown Sampling

Density of Gumbel variables with location parameter $\log \alpha$ and scale 1:

$$f_{\log \alpha}(g) = \exp(-g + \log \alpha) \exp(-\exp(-g + \log \alpha)) = \exp(-g + \log \alpha)F_{\log \alpha}(g).$$

Let $K$ be the number of discrete states, and $F_{\log \alpha}(g)$ is the CDF of the Gumbel variable with location $\log \alpha$. Without any prior on the target Gumbels, the location parameters in the target domain can simply be obtained according to the transition probabilities estimated naively from limited data. That is $p(\alpha) = \delta(\log F'(\cdot))$ where $\delta$ is the dirac-delta distribution. Sampling from this Gumbel given observation $k'$ can be done using Topdown procedure [Oberst and Sontag 2019]. [Madison et al. 2014].

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

\[
p(k', g_{1}^{(T)}, \ldots, g_{n}^{(T)}) = \\
\int_{\alpha} \frac{\alpha^{j}}{Z} \prod_{i \in X} \left[ f_{\log \alpha_{i}}(g_{i}^{(T)}) | g_{j}^{(T)} \geq g_{i}^{(T)} \right] p(\alpha) d\alpha
\]

Equation 6 can be obtained exactly following Maddison et al. [2014]. That is, for a fixed and known $\alpha$, they show that in the posterior, the gumbel corresponding to the observed outcome $k'$, i.e. $g_{k'}^{(T)}$ is a Gumbel variable with location parameter $Z = \log \sum_{i=1}^{K} \alpha_{i}$, the max value $k'$ and Gumbels are independent and the rest of the exogenous variables $g_{i}^{(T)}|i \neq k'$ are truncated by the max gumbel value corresponding to $k'$ (shown by the Iverson brackets above). The dirac-delta prior is now replaced by the mixture prior from Equation 2.

The sampling procedure follows a modified top-down procedure such that for every counterfactual sample, we first select the mixture component with probability $[u^{(T)}, 1 - u^{(T)}]$, followed by posterior sampling over the Gumbels.

S1.4 Mixture-prior preserves counterfactual stability

Definition S1.4. Counterfactual Stability: An SCM over discrete random variables is counterfactually stable if: If we observe $X_{i} = i$, then $\forall j \neq i$, if $\frac{P_{i}}{P_{j}} \geq \frac{P_{i}}{P_{j}}'$, implies that $P^{do(\tau') | X_{i}=i}(X = j) = 0$.

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 distributions over $\pi_{i}$ for which expected rewards are to be estimated and $\pi_{j}$ for which expected rewards are to be estimated and $\pi_{j}$ for which expected rewards are to be estimated and $\pi_{j}$ for which expected rewards are to be estimated.

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: $\mu^{(T)}$ be the behavior policy in the target environment and the corresponding trajectories denoted by $\tau^{(T)}$. Let $\pi^{(T)}$ be a candidate policy for which expected rewards are to be estimated and $\pi^{(T)}$ be the counterfactual trajectories using conditional posteriors $p(U^{(T)} | \tau)$ over exogenous variables $U^{(T)}$, $\tau$ is a deterministic function of $U^{(T)}$. The prior distributions over $U$ are $p^{(T)}(U^{(T)}) = p(U^{(T)}) = p(U^{(T)})$ (which remains the same as any informative prior coming from the source environment imposed in this framework). We drop the notation (T) in the following as we are only concerned about the target environment hereon. Source distributions, if any, will be

https://csadis.github.io/gumbel-machinery
We consider two subpopulation groups (diabetic) and (non-diabetic) and the corresponding transition dynamics. 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 $g^{(T)} \subseteq \mu^{(T)}$, the posterior is inferred by incorporating the mixture prior that helps regularize from the source.

### S2.1 Justification of Counterfactual Regularization

We consider two subpopulation groups (diabetic) and (non-diabetic) and the corresponding transition dynamics $P_d(S,A)$ and $P_{nd}(S,A)$. 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 $\hat{P}_d^{(B)}(S,A)$ and $P_d^{(T)}(S,A)$ is higher in the source domain than in the target domain (by virtue of number of samples from this sub-population 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:

$$\| \hat{P}_d^{(B)}(S,A) - P_d(S,A) \| \geq \| P_d^{(T)}(S,A) - P_d(S,A) \|$$

$$\| P_{nd}^{(B)}(S,A) - P_{nd}(S,A) \| \leq \| P_{nd}^{(T)}(S,A) - P_{nd}(S,A) \|$$

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): $\eta \hat{P}_d^{(B)}(S,A) + (1-\eta) \hat{P}_d^{(T)}(S,A)$ where $0 \leq \eta \leq 1$.

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

$$\| P_{nd}(S,A) - \eta \hat{P}_{nd}^{(B)}(S,A) + (1-\eta) \hat{P}_{nd}^{(T)}(S,A) \| \leq \| P_{nd}(S,A) - \hat{P}_{nd}^{(B)}(S,A) \| + (1-\eta) \| P_{nd}(S,A) - \hat{P}_{nd}^{(T)}(S,A) \|$$

$$\| P_{nd}^{(B)}(S,A) - P_{nd}(S,A) \| \leq \| P_{nd}(S,A) - \hat{P}_{nd}^{(T)}(S,A) \|$$

$$\| \hat{P}_{nd}^{(B)}(S,A) - P_{nd}(S,A) \| \leq \| \hat{P}_{nd}^{(T)}(S,A) - P_{nd}(S,A) \|$$

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 $\log \hat{P}_d^{(B)}(S,A)$ and $\log \hat{P}_d^{(T)}(S,A)$ 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 sub-population.
S3 KL-aggregation for CF-PI

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

\[ \pi = \arg\min_{\pi} \text{AKL}(\pi \parallel \nu) + (1 - \lambda) \text{KL}(\pi \parallel \pi^{(S)}) \]  

(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:

\[ \arg\min_{\pi} \lambda \langle \pi^T, \log \pi - \log \nu \rangle + \langle \pi^T, \log \pi - \log \pi^S \rangle \]
\[ \text{s. t. } \pi \in \Delta^{K-1} \]  

(20)

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

\[ \arg\min_{\pi} \lambda \langle \pi^T, \log \pi - \log \nu \rangle + \langle \pi^T, \log \pi - \log \pi^S \rangle + \mu \left( \sum_{k=1}^{K} \pi_k - 1 \right) + \beta \pi \]

where \( \beta \geq 0 \)

Taking the gradient and setting to 0 yields:

\[ (1 + \log \pi) + \mu 1 + \beta = \lambda \log \nu + (1 - \lambda) \log \pi^S \]

(22)

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

S4 CFPT Procedure

Algorithm 3 Counterfactually Guided Policy Transfer

1: // Counterfactual inference (CFI) with source environment prior
2: CFI(data \( \bar{x}_s \), SCM \( \mathcal{M}_s \), intervention \( I \), query \( X_q \), prior \( \pi^{(S)} \))
3: \( \bar{u} \sim p(u | \bar{x}_s) \) \{Sample noise variables from posterior over latent parameters\}
4: \( p(u) \leftarrow \delta(u - \bar{u}) \) \{Replace noise distribution in \( p \) with \( \bar{u} \)\}
5: \( f_i \leftarrow f_i' \) \{Perform intervention \( I \)\}
6: return \( Z \sim p \left( \eta | \pi^{(S)}(x_i) | \bar{u} \right) \) \{Simulate from the counterfactual posterior over model \( \mathcal{M}_s \), Alg. 3\}

7: // Regularized Policy Iteration (RegPI)
8: RegPI(current policy \( \pi^{(T)} \), discount \( \gamma \), aug. statistics \( \tilde{P}^{(T)} \), source policy \( \pi^{(S)} \), reg. param \( \lambda \))
9: Initialize \( V(s) \) for all \( s \in S \)
10: repeat
11: repeat
12: for each \( s \in S \) do
13: \( \nu \leftarrow V(s) \)
14: \( V(s) \leftarrow \sum_{s'} \tilde{P}^{(T)}(s'|s, \pi^{(T)}(s)) \left[ R(s, \pi^{(T)}(s)) + \gamma V(s') \right] \)
15: end for
16: until convergence
17: for each \( s \in S \) do
18: \( \nu(s) \leftarrow \frac{1}{Z} \left[ R(s, \nu) + \gamma \left( \tilde{P}^{(T)}(s'|s, \nu) \otimes V(s') \right) \right] \) \{Gen. a proposal dist. over actions\}
19: \( \pi^{(T)}(s) \leftarrow \arg\max_{\pi} \exp \left( \lambda \log \nu(s) + (1 - \lambda) \log \pi^{(S)}(a|s) \right) \) \{KL minimization, Eq. 22\}
20: end for
21: until \( \pi^{(T)} \) converges or after MAX_ITERATIONS
22: // Counterfactual Policy Iteration (CFPI)
23: CFPI(SCM \( \mathcal{M}_s \), init. policy \( \pi^{(T)}_0 \), source policy \( \pi^{(S)} \), source statistics \( \tilde{P}^{(S)} \), num. iters \( K \), num. traj samples \( N \), mixture param \( \eta \))
24: for \( k = 1, \ldots, K \) do
25: // Gather a batch of counterfactually generated trajectories in the target environment
26: \( \{h^{(k)}_i\}_{i=1}^{N} \sim \mathcal{H}^{(T)} \subset \mathcal{T} \) \{Sample batch of trajectories from observed data\}
27: \( \{\tau^{(T)}_i\}_{i=1}^{N} = \text{CFI}(\{h^{(k)}_i\}_{i=1}^{N}, \mathcal{M}, I(\mu \rightarrow \pi^{(T)}_k), \mathcal{T}, \tilde{P}^{(S)}) \) \{Counterfactual rolls under \( \pi^{(T)}_k \)\}
28: // Estimate empirical transition statistics \( \tilde{P}^{(T)} \) from \( \{\tau^{(T)}_i\}_{i=1}^{N} \)
29: \( \tilde{P}^{(T)} = \frac{1}{N} \eta \tilde{P}^{(T)} + (1 - \eta) \tilde{P}^{(S)} \) \{Augment observed environment transition statistics\}
30: \( \pi^{(T)}_k \leftarrow \text{RegPI}(\pi^{(T)}_k, \gamma, \tilde{P}^{(S)}, \lambda) \)
31: end for

Here we present the pseudocode (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 \( \pi^{(T)} \) developed within a data-rich source environment \( S \) as well as an estimation of the transition function.
statistics $P^{(S)}$ collected from observed data. These methods combine to form a two-phase counterfactual regularization approach for policy learning in a data-scarce target environment $\tau$.

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 in the target environment are sampled (Alg. 3 line 24). This batch is used, along with the current policy within $\tau$, $\pi_k^{(\tau)}$, and the prior over the transition statistics from the source environment $P^{(S)}$ to generate counterfactual trajectories $\{\tau^{(i)}\}_{i=1}^N$ (Alg. 3 line 25 → CF in lines 1-6). This counterfactual sampling procedure, leveraging 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 $\tau$. To account for this, an augmented transition matrix $P^{(T)}$ is formed through a weighted sum between $P^{(S)}$ and the empirically observed set from $\{\tau^{(i)}\}_{i=1}^N$ ($P^{(T)}$, line 20). This augmented transition matrix is then passed to RegPI as discussed in Section 4.2 (line 27 → RegPI, lines 7-21).

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

S5 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.

S5.1 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](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 $\gamma$ 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 $\mathcal{S}$ was fixed to 10,000 and the proportion of diabetic patients in $\mathcal{S}$ was set to 0.1. All target environments $\tau$, 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 `gpytoolbox` library. Code to replicate our experiments will be made available upon publication of our paper.

S5.1.1 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 $\mathcal{H}^{(T)}$ which is then used within policy iteration to produce the policy $\pi^{(T)}$.

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

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

S5.1.2 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 $\tau$ when transferring from a fixed source environment (proportion of diabetic patients: 0.1). This involved
What we see in Figure S1 is that there is a balance when selecting $\mu$.

All other procedures and operations within CFPT were run as normal with the same parameter settings as using the Gumbel-Max SCM, regularized by $\eta$ and $\lambda$.

In S5.2.2 Counterfactual sampling with fully observed state patients in $T$ are the values used for all CFPT variants and ablations presented in Sec 5 when the proportion of diabetic patients observed in the target clinical environment. In Table S2 we present the numerical values for the transitions completely from CFPT. We also removed any batch sampling from $P(T)$.

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 S1 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 $H(T)$, using instead the full set of observed data within $T$. A single run of RegPI was executed, using the top performing values for $\lambda$ as reported in Table S1.

S5.2 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 S2 we present the numerical values for the comparison between CFPT and all baselines and ablations shown in Figure S1.

| Diabetic Proportion | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $w(T)$             | 0.8 | 0.8 | 0.8 | 0.6 | 0.7 | 0.8 | 0.8 | 0.6 | 0.8 | 0.7 | 0.8 |
| $\eta$             | 0.7 | 0.8 | 0.7 | 0.7 | 0.8 | 0.7 | 0.6 | 0.8 | 0.7 | 0.7 | 0.7 |
| $\lambda$          | 0.9 | 0.9 | 0.3 | 0.1 | 0.3 | 0.6 | 0.3 | 0.1 | 0.3 | 0.4 | 0.9 |

S5.1.3 Ablations

Determining the best value for the number of iterations $K$ of CF-PI, the mixture weight for regularizing the counterfactual sampling $w(T)$, the weighting for augmenting the observed transition statistics $\eta$, and perhaps most importantly the weight for regularizing the policy learning with $\lambda$. As $w(T)$, $\eta$ and $\lambda$ correspond to linear combinations between two quantities, we tested each of these hyperparameters between 0 and 1 in increments of 0.1, using the learned policy’s true RL performance in the target environment to compare between settings. We report the optimal settings for learning within $T$ in each target environment (diabetic proportion of population ranging from 0 to 1 in 0.1 increments) in Table S1. For all target environments the number of iterations $K$ for CF-PI was 50.

S5.2.1 Analysis of selecting $\eta$, affecting the augmentation of $P(T)$

In Figure S1 we demonstrate the range of policy performance under CFPT with CF-PI when varying the parameter $\eta$. Recall from Section 4.3 that $\eta$ is used to weight the augmentation of the observed transition statistics in the target environment ($P(T)$) with those estimated from the counterfactually inferred trajectories ($P(T)$). 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 $\lambda$ and the parameter $\eta$ which is used to incorporate the inferred empirical transition matrix $P(T)$ into the observed target transition matrix $P(T)$ for use in regularized policy iteration (Algorithm S1, line 26).

What we see in Figure S1 is that there is a balance when selecting $\eta$ and $\lambda$ for CFPT policy learning. As $\lambda$ increases, meaning we are using less of the source environment, no matter the choice of $\eta$, performance more or less converges to the baseline non-transfer setting within $T$. However when $\lambda$ is smaller, meaning we intend to use a larger proportion of the source policy, we see that the choice of $\eta$ can have a broad effect. In the scenario demonstrated in Figure S1 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 $T$ is 0.8.

S5.2.2 Counterfactual sampling with fully observed state

Similar to the analysis presented in Section 4.3 and in Figure 6, we investigate the change in the feature distributions in $T$ when the simulated patient’s diabetic status is known after sampling counterfactual trajectories using the Gumbel-Max SCM, regularized by $z$. The resulting comparison is shown in Figure S2.

| Approach | True RL Reward | WIS Reward |
|----------|----------------|------------|
| Scratch  | $-0.7398 \pm 0.007$ | 0.6388 $\pm$ 0.584 |
| Pooled   | $-0.4808 \pm 0.012$ | 0.9782 $\pm$ 0.004 |
| Blind    | $-0.3915 \pm 0.013$ | 0.5874 $\pm$ 0.113 |
| RegPI    | $-0.3566 \pm 0.012$ | 0.6266 $\pm$ 0.057 |
| Red. CFPT| $-0.2116 \pm 0.010$ | 0.7689 $\pm$ 0.077 |
| CFPT     | $-0.1491 \pm 0.011$ | 0.7333 $\pm$ 0.044 |

S2: Numerical values corresponding the policy performance results presented in Figure 6. The observed behavior policy $\mu(T)$ receives an average reward of 0.1486 $\pm$ 0.018.
S1: 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 \( \lambda \) and \( \eta \) interact with a diminished effect as \( \lambda \) increases. CF-PE estimated reward (dotted lines) asymptotically overestimates policy performance as \( \lambda \) increases.

S2: Feature distributions with full observations with the patient observations obtained in \( T \) on the left and a resampling of the feature distributions using counterfactuals drawn from the regularized Gumbel-Max SCM on the right.

S5.2.3 Off-policy Evaluation of \( \pi(T) \)

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. S5.3.3 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 \( \pi(T) \) in Table S2 for the setting of \( T \) with a 0.8 proportion of diabetic patient trajectories. As expected, WIS overestimates the true RL return in \( T \), 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. S5.3.3, CFPT acts more conservatively and closely approximates the observed behavior, leading to more stable performance.

S5.3 Qualitative Analysis of \( \pi(T) \)

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 $\tau$ (other comparisons between CFPT and the baselines are in Section S5.3.3). 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 $\tau$. In Section S5.3.2 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 S4, 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.

S4: 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 S5.3.3.

S5.3.1 Sub-population Analysis of Evaluated Policies

In Figure S5 we demonstrate the differences among subpopulations when learning a policy with CFPT for different target environments $\tau$ (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 $p_{Diab} = 0.1$. The population of this source environment is distributionally similar to $T$ when $p_{Diab} = 0.2$. 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 $p_{Diab}$ settings in Figure S5. This provides some further evidence toward our conclusion that CFPT aids in the development of more circumspect policies.

In Figure S6 we demonstrate the differences among subpopulations when learning a policy with CFPT having different settings of $\eta$ (see Section S5.2.1). With a properly chosen $\eta$ (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.

In Figure S7 we present 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.

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 S4, we include four additional trajectory introspection figures. The first of which belongs to a non-diabetic patient (Figure S9 recall, this is type of patient is found in lower proportion within the target environment) while the other three are diabetic patients (Figures S10, S12).
S7: 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.

S8: Comparison of all baselines in their aggregate population statistics in counterfactual evaluation of the policies learned in the target environment pDiab=0.8.
S9: 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.

S10: 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.
S11: 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.

S12: 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.