A Appendix

A.1 Bayesian RL for Dynamics and Rewards Estimation.

We assume that the dynamics are governed by a sequence of MDPs to allow for non-stationarity. The dynamics of each MDP in the sequence, denoted by $\mathcal{P}_t$ is estimated using batch data. Our batch data $\mathcal{D}^*$ can be split into tuples indexed by the time t as: $\mathcal{D}^* \triangleq \{\mathcal{D}_t^*\}_{t=1}^T$. To infer $\mathcal{P}_t$, batch samples $\mathcal{D}_t^*$ will be used.

Discrete State. We first describe the procedure for discrete states. The distribution characterizing the dynamics $\mathcal{P}_t$ are parameterized by $\theta_t \triangleq \{\theta_t(s,a)\}_{(s,a)}$ which is a tuple over all state-action pairs. In particular, we use a Bayesian framework to estimate the parameters of the dynamics distribution $\mathcal{P}_{\theta,t}$. For discrete data, $\mathcal{P}_{\theta,t} \triangleq p(\mathcal{D}_t^*|\theta_t)$ are modeled as Multinomial distributions for each state-action pair (s,a). That is, this parameter $\theta(s,a) \in \Delta^{|\mathcal{S}|-1}$ lies in the probability simplex of dimension $|\mathcal{S}|-1$ and $p(s'|s,a) = \theta_{s'}(s,a)$. The prior $p(\theta_t(s,a))$ is assumed to follow a Dirichlet Distribution (Fink, 1997). The Dirichlet prior distribution is parameterized by $\alpha \in \mathbb{R}^{|\mathcal{S}|-1}$ and is given by,

$$p(\theta) = \frac{1}{B(\alpha)} \prod_{i=1}^{|\mathcal{S}|} \theta_i^{\alpha_i - 1}$$

where $B(\alpha)$ is the multivariate Beta function.

The posterior distribution of $\theta(s,a)$ given samples $\mathcal{D}_t^*$ is itself a Dirichlet distributed random variable (since Dirichlet distribution is a conjugate prior for the Multinomial likelihood distribution), with parameters α_t',

$$p(\theta_t(s,a)|\mathcal{D}_t^*) = \text{Dirichlet}(\alpha_t') \tag{4}$$

where $\alpha'_{s',t} = \alpha_{s',t} + \sum_{s'' \in \mathcal{D}_t^*} \mathbf{1}(s'' == s')$.

Note that we also account for uncertainty over rewards by estimating posteriors via Bayesian inference, as in the case of the dynamics. For discrete rewards, Dirichlet priors are used analogously.

Continuous State. For continuous states, the dynamics are assumed normally distributed. The parameters of the normal distribution are assumed to have a normal-gamma prior (Fink, 1997). That is, Let $\mu_t(s,a), \tau_t(s,a)$ be the mean, and precision of the parameter describing the dynamics as a function of the states and actions. The normal-gamma prior is given as follows:

$$\begin{aligned} \mu_t(s,a)|\tau(s,a) &\sim \mathcal{N}(\mu_0, n_0\tau(s,a)) \\ \tau(s,a) &\sim \text{Gamma}(\eta, \beta) \end{aligned} \tag{5}$$

Note that we use the same prior distribution for all state-action pairs, though custom priors may be used based on domain-knowledge. The posterior distributions after observing data samples $\mathcal{D}_t^*$, specifically $\mathbf{s}'$ over $\mu_t(s,a)$ are given by a Gaussian with the following parameters:

$$\mu(s,a)|\tau(s,a), \mathbf{s}' \sim \mathcal{N}\Big(\mu_t', \tau_t'\Big) \tag{6}$$

$$\mu_t' = \frac{n_t\tau(s,a)}{n_t\tau(s,a) + n_0\tau(s,a)}\bar{s'} + \frac{n_0\tau(s,a)}{n_t\tau(s,a) + n_0\tau(s,a)}\mu_0 \tag{7}$$

$$\tau_t' = n_t\tau(s,a) + n_0\tau(s,a) \tag{8}$$

where $\bar{s'}$ is the mean of the observations $\mathbf{s}'$ and n_t are the number of observations for state-action s,a observed at time t. The posterior distribution over the precision $\tau(s,a)$ is given by,

$$\tau(s,a)|\mathbf{s}' \sim \text{Gamma}(\eta', \beta')$$

where,

$$\eta' = \eta_0 + \frac{n_t}{2}, \beta' = \beta_0 + \frac{1}{2}\left(n_t \sum_i (s'_i - \bar{s})^2 + \frac{n_t n_0 (\bar{s}' - \mu_0)^2}{2(n_t + n_0)}\right)$$

A note on choice of conjugate priors: Conjugate priors usually (though not always) map one-to-one to the likelihood distribution. That is, the Dirichlet prior is the only conjugate prior for the multinomial distribution (which we use for the discrete case) and the Normal-Inverse Gamma prior is a very common conjugate prior for the Normal distribution (which we use in the continuous case). The entire list of conjugate priors is available in Fink (1997).

A.2 Decomposing Propagated Uncertainty

We describe how uncertainty of the long-term outcome $\mathbb{E}[r_T|s_{t_d}, , \mu_{t_d}]$ at deferral time t_d, when the agent is in state s_{t_d} can be decomposed into modeling/epistemic uncertainty and irreducible/aleatoric uncertainty in the following. Once we defer, we sample actions from π_0 at time $t' = t_d$ and π_{mix} for $t' > t_d$ where the mixture probability is determined by $g_{\pi_{\text{tar}}}$ for future deferrals. The expected long-term outcome is given by:

$$\mathbb{E}[r_T|s_{t_d}, \mu_{t_d}] = \int_{s_{t_d+1}}^{s_T} \int_{a_{t_d}}^{a_T} \int_{\mu_{t_d+1}}^{\mu_T} \int_{\theta_{t_d}}^{T} r(s_T, a_T) \times \prod_{t'=t_d+1}^{T} p_{t'}(s_{t'}|\mu_{t'}) p_{t'}(\mu_{t'}|\theta'_t(s_{t'}, a_{t'})) \pi_{t'}(a_{t'}|s_{t'}) p_{t'}(\theta_{t'}|\mathcal{D}) d\mathbf{s} d\mathbf{a} d\boldsymbol{\mu} d\boldsymbol{\theta}$$

Integrands are written in short-hand: $\mathbf{s} = \{s_{t_d+1}, s_{t_d+2}, \cdots, s_T\}$ (analogously for other quantities). As suggested before, $\pi_{t_d} = \pi_{0(t_d),\text{mix}(t_d+)}$ to account for future deferrals. Here, we denote the posterior MDP samples for any state action pair by μ_t. The variability in these samples capture modeling uncertainty. The dynamics parameters are denoted by θ_t for each state-action pair. We sample from posterior distribution $p(\theta_{t'}|\mathcal{D}^*)$, followed by sampling the MDPs $\mu_t \sim p(\mu_{t'}|\theta'_t(s_{t'}, a_{t'}))$. We maintain one estimate of parameter $\theta_{t'}$ and sample K MDPs $\mu_{t'}$ from this distribution. That is, $p(\theta_{t'}|\mathcal{D}) = \delta_{\theta_{t'}}$ which is a delta function centered at $\theta_{t'}\, \forall\, t' \in \{0, 1, 2, \cdots, T\}$:

$$\mathbb{E}[r_T|s_{t_d}, \mu_{t_d}] = \int_{s_{t_d+1}}^{s_T} \int_{a_{t_d}}^{a_T} \int_{\mu_{t_d+1}}^{\mu_T} \int_{\theta_{t_d}}^{T} r(s_T, a_T) \times \prod_{t'=t_d+1}^{T} p_{t'}(s_{t'}|\mu_{t'}) p_{t'}(\mu_{t'}|\theta'_t(s_{t'}, a_{t'})) \pi_{t'}(a_{t'}|s_{t'}) \delta_{\theta_{t'}} d\mathbf{s} d\mathbf{a} d\boldsymbol{\mu}$$

Thus, the epistemic uncertainty we capture is due to the uncertainty over dynamics under fixed parameters. High variability in sampling $\mu_{t'}$ indicate the current state $s_{t'}$ (and action) are out-of-distribution. The total uncertainty can now be decomposed using the law of total variance:

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$$

Applying this to our target outcome, we have:

$$\underbrace{\text{Var}(r_T|s_{t_d}, \mathcal{D})}_{\text{Total Uncertainty}} = \underbrace{\mathbb{E}_{\mu_{t_d} \sim p(\mu_{t_d}|\mathcal{D})}[\text{Var}(r_T|\mu_{t_d}, s_{t_d}, \mathcal{D})]}_{\text{Irreducible/ Aleatoric Uncertainty}} + \underbrace{\text{Var}_{\mu_{t_d} \sim p(\mu_{t_d}|\mathcal{D})}(\mathbb{E}[r_T|\mu_{t_d}, s_{t_d}, \mathcal{D}])}_{\text{Epistemic/Modeling Uncertainty}} \tag{9}$$

The second term is the variance *conditioned* on knowledge of the model μ_{t_d}, therefore marginalizing only over current aleatoric uncertainty and future total uncertainty (i.e. over future $\mu_{t'}$, future deferral, and reward). This is the *propagated uncertainty due to modeling uncertainty at* t_d, and can be reduced by data collection. The first term averages over the variance due to μ_{t_d} and captures *propagated uncertainty to due to aleatoric uncertainty at* t_d, which can only be reduced by careful interventions at t_d. We estimate these using Monte-Carlo sampling. High *propagated epistemic uncertainty* conveys that the current uncertainty of model prediction (of the dynamics) is high but could be improved if additional data could be collected. High *propagated aleatoric uncertainty* indicates high variability in the patient's dynamics that can only be with careful interventions and is otherwise not manageable. Based on the communicated uncertainty, the clinician may choose to deviate from their usual practice for rare cases with high epistemic uncertainty and instead consult multiple experts or attempt experimental treatments.

A.3 Datasets

Discrete Toy Data. All state and action spaces are discrete. True dynamics of the environment are known. This environment has 8 discrete states and binary actions $\{a_0, a_1\}$. All samples start at state 0 and progress toward a sink state 7. The episode length is 15. State 6 has low reward (-5) while all other states have a reward of $+1$. The initial dynamics are set up such that action a_0 reduces the probability of landing in stage 6, and action a_1 increases the probability of reaching state 6. π_{tar} increases the chances to reach state 6 unfavourably by taking action a_1 in states $2, 3, 4$ when $t < 5$ or $t > 12$. We expect to defer in states $2, 3, 4$ even though rewards are favorable, if a method is pre-emptive. When $5 \leq t \leq 12$, the dynamics flip such that a_0 becomes an unfavourable action that increases the probability of landing in 6, while a_1 reduces this probability. Here, π_{tar} again increases the chances of landing in 6, by taking a_0 more often in states $2, 3, 4$. By flipping the better action to a_0 in this region, it becomes crucial to *estimate the dynamics* over predicting the best action. The dynamics are non-stationary and the probability of landing in state 6 progressively increases when $5 \leq t \leq 12$. The reward vector for Discrete Toy data is a function of states only and is given by:

$$r(s) = \begin{cases} 1 & \text{if } s \neq 6 \\ -5 & \text{if } s = 6 \end{cases}$$

We collect $N = 500$ episodes for simulated data.

Diabetes simulator. We use an open-source implementation of the FDA approved Type-1 Diabetes Mellitus simulator (T1DMS) for modelling treatment of Type-1 diabetes. The simulator models the dynamics of an in-silico patient's blood glucose levels when consuming a meal. If the blood glucose level is either too high (hyperglycemia) or too low (hypoglycemia), this can have fatal consequences such as organ failure. As a result, a clinician must administer an insulin dosage to minimize the risk of such events. While a doctor's initial dosage prescription is usually available, the insulin sensitivity of a patient's internal organs changes over time, thereby introducing non-stationarity that should be accounted for. We sample 10 adolescent patient trajectories (episodes) over 24 hours (with measurements aggregated at 15 minute intervals). Glucose levels are discretized into 13 states according to ranges suggested in the simulator. Further, insulin and bolus intervention combinations are discretized to generate a total of 25 actions. The non-stationary characteristics of this publicly available simulator allow us to thoroughly evaluate all aspects of SLTD under the true dynamics of the simulator. The reward function for Diabetes data is stationary and defined in the simulator. It is a function of the state and is defined as the change in risk due to change in blood glucose level between the last two measurements. Discretization of Glucose levels is provided in Table 3 and discretization of interventions is summarized in Table 4 (bolus) and Table 5 (insulin). The discrete combinations of bolus $\{b0, b1, b2, b3, b4\}$ and insulin $\{i0, i1, i2, i3, i4\}$ are combined to generate 25 potential actions: $\{b0 - i0, b0 - i1, b0 - i2, b0 - i3, \cdots, b4 - i4\}$.

We introduce non-stationarity within each episode by increasingly changing the adolescent patient properties to an alternative patient over the episode. This is different from the setting of Chandak et al. (2020b) where non-stationarity is indeed across episodes. Thus our setting is more challenging. This significantly affects the utility of the initial target policy necessitating deferral as the patient properties change over the course of the day. The non-stationary clinician/behavior policy is estimated using Q-learning. We use an epsilon-greedy version of such a policy.

Incorporating non-stationarity into the simulator: We use the "Navigator" sensor to generate blood-glucose measurements and the "Insulet" pump to simulate interventions. For each episode, non-stationarity is induced by modifying the patient configurations over a period of 24 hours. This result in different dynamics over the course of the day. These configurations modify insulin sensitivity, glucose absorption and the insulin action on glucose production among other parameters. For each episode, two random adolescent patients are sampled (say 'a', and 'b'), over every minute the patient parameters are then sampled as a convex combination of patient 'a' and patient 'b' where, as we progress in time, the convex combination increasingly shifts from 0 to 1 thus changing patient parameters. Over the episode, the patient parameters increasingly look like that of patient 'b' instead of 'a'. The rate of change of this convex combination can be controlled and is set to $\cos(t \times \text{speed} \times 0.0005) \times 0.5 + 0.5$, where speed $= 5$ for our simulations. A similar policy was used by Chandak

Blood Glucose (mg/dL) - BG	Discrete state
$0 < \text{BG} \leq 29.2$	0
$29.2 < \text{BG} \leq 58.5$	1
$58.5 < \text{BG} \leq 87.8$	2
$87.8 < \text{BG} \leq 116.9$	3
$116.9 < \text{BG} \leq 146.2$	4
$146.2 < \text{BG} \leq 175.4$	5
$175.4 < \text{BG} \leq 204.6$	6
$204.6 < \text{BG} \leq 233.9$	7
$233.9 < \text{BG} \leq 263.1$	8
$263.1 < \text{BG} \leq 292.3$	9
$292.3 < \text{BG} \leq 321.6$	10
$350.8 < \text{BG} \leq 365.4$	11
> 365.4	12

Table 3: Discretization of Blood Glucose for the Diabetes simulator

Bolus (g/min)	Action
0.00 - 18.6	$b0$
18.6 - 37.2	$b1$
37.2 - 55.8	$b2$
55.8 - 74.4	$b3$
> 74.4	$b4$

Table 4: Discretization of bolus for creating combination treatments

Insulin (U/min)	Action
0.00 - 2.5	$i0$
2.5 - 5.5	$i1$
5.5 - 8.5	$i2$
8.5 - 11.5	$i3$
11.5 -	$i4$

Table 5: Discretization of insulin for creating combination treatments

State Variables
$CD4^+$ count
$CD8^+$ count
Viral Load
Glycoprotein Mutations GP41101V - GP41296A
Protease Mutations PR10I - PR98S
Reverse Transcriptase Mutations RT64K - RT396D
Integrase Mutations INT10E - INT721

Table 6: State Variables considered in HIV Case I and II

et al. (2020b) to induce non-stationarity. However Chandak et al. (2020b) do not induce non-stationarity within an episode, but across different episodes. The target policy π_{tar} is learned on data collected from patients whose dynamics do not change over time.

HIV Data. This dataset is publicly available upon request and is a continuous state dataset. We identified individuals between 18-72 years of age from the EuResist database (Zazzi et al., 2012) comprising of genotype, phenotype and clinical information of over 65,000 individuals in response to antiretroviral therapy administered between 1983-2018. Patients are administered combinations of different drugs to prevent drug resistance and the development of viral mutations that could potentially result in resistance. Once resistance to a particular drug occurs, it is also possible for cross-resistance to develop to similar antiretrovirals from the same class, thus limiting a patient's potential treatment options for the future. As a result, a clinician must administer antiretrovirals to minimize the risk of such resistance, while lowering the viral load in the blood. While several therapy guidelines based on clinician expertise are available, depending on what therapies a patient has previously been administered, several new mutations may develop in response to therapy, resulting in new drug-resistant variants to emerge that potentially change over time, thereby introducing non-stationarity that should be accounted for. Viral evolution (via the development of mutations and resistance to certain drugs) across different populations has led to the emergence of different HIV strains, some of which are easier to treat than others.

We focus on $32,960$ patients' genotype, treatment response, CD4+ and viral load measurements, gender, age, risk group, number of past treatments collected over 14 years (aggregated at 4-6 month intervals) producing trajectories of on average $T = 16$ steps. Our state space consists of continuous states of cell counts, viral loads and mutations. Names of the variables comprising the states are provided in Table 6.

Drug combinations are discretized to produce 25 actions of the most frequently occurring combinations. The reward for HIV is function of the viral load (L_t), the immune response (C_t) and the mutations (M_t) and is given by

$$r(s_t) = \begin{cases} -0.7 \log L_t + 0.6 \log C_t - 0.2 M_t, & \text{if } L_t \text{ is above 40 copies/mL} \\ 5 + 0.6 \log C_t - 0.2 M_t & \text{if } L_t \text{ is below 40 copies/mL} \end{cases} \tag{10}$$

Further details motivating the choice of this reward can be found in Parbhoo et al. (2017; 2018).

For our first case study, we investigate whether deferring to a second line therapy (Saag et al., 2020) in response to potential drug resistance improves long-term outcomes. Here, the clinician policy corresponds to a second line therapy (provided by our clinical collaborators), while the non-stationary behaviour policy corresponds to a first line therapy estimated using Q learning.

For our second case study, the non-stationary behaviour policy corresponds to a first line therapy typically used for treating patients of subtype C. We then examine whether deferring to a first line therapy (provided by our clinical collaborators) used for treating patients of subtypes A, B, D, E, F, G (due to potential drug resistance) improves long-term outcomes. The therapies considered for both HIV Case I and II are shown in Table 7.

Drug Combination	Discrete Action
$TDF + FTC + EFV$	0
$TDF + FTC + ATV/r$	1
$3TC + AZT + LPV/r$	2
$TDF + FTC + DRV/r$	3
$TDF + FTC + LPV/r$	4
$3TC + ABC + LPV/r$	5
$3TC + TDF + d4T$	6
$3TC + AZT + EFV$	7
$3TC + ABC + EFV$	8
$3TC + TDF + LPV/r$	9
$3TC + AZT + NVP$	10
$3TC + ddl + EFV$	11
$TDF + FTC$	12
$3TC + d4T + LPV/r$	13
$3TC + DRV/r + ABC$	14
$EFV + 3TC$	15
$EFV + 3TC + FTC$	16
$DTG/3TC/TDF$	17
$DTG/3TC$	18
$AZT + TDF + LPV/r$	19
$d4T + NFV$	20
$ddl + d4T + SQV/r$	21
$ddl + TDF + EFV$	22
$AZT + ddl + NVP$	23
$ddl + d4T + IDV/r$	24

Table 7: Drug combinations considered for HIV Case I and II

B Additional Results

	Synthetic data			Diabetes		
Method	Self-Normalized IS (mean ± 2 s.e.)	(Best) Cost hyperparameter	Deferral Frequency (True Dynamics)	Self-Normalized IS (mean ± 2 s.e.)	(Best) Cost hyperparameter	Deferral Frequency (True Dynamics)
SLTD	3.710 ± 0.078	0.00	0.220	-1.76e+05 ± 1.00e+04	0.00	0.515
SLTD-Stat.	4.170 ± 0.202	0.00	0.255	-2.27e+06 ± 7.30e+03	0.00	0.508
SLTD-One Step	-1.283 ± 0.015	0.02	0.005	-1.16e+05 ± 9.97e+03	0.00	0.518
SLTD (K=1)	3.819 ± 0.125	0.02	0.204	-1.87e+05 ± 6.55e+04	0.00	0.514
SLTD-Stat. (K=1)	4.028 ± 0.265	0.00	0.237	-2.17e+06 ± 1.08e+05	0.00	0.520
SLTD-One Step (K=1)	-1.170 ± 0.081	0.05	0.007	-9.46e+04 ± 4.81e+04	0.00	0.518
Augmented-MDP	2.731 ± 0.00	0.00	0.510	-2.64e+06 ± 0.000	0.00	0.000
Mozannar et. al.	4.111 ± 0.073	0.20	0.912	-2.37e+06 ± 1.39e+04	0.20	0.073
Madras et. al.	4.743 ± 0.083	0.01	0.355	-8.148 ± 0.026	0.02	1.00
π_{tar}	-1.397 ± 0.00	N/A	N/A	-2.44e+06 ± 0.000	N/A	N/A
π_0	11.564 ± 0.00	N/A	N/A	-8.085 ± 0.000	N/A	N/A

Table 8: Expected Value using Self-Normalized IS for SLTD and all baselines with corresponding cost hyperparameters and deferral frequency. As we can see, for Synthetic data, SLTD achieves the best tradeoff while most other baselines are only able to achieve this value for 0 cost for SLTD. Overall there are biases in the IS estimate compared to the values from true dynamics. For Diabetes , we see that support challenges make it difficult for IS to distinguish clearly between different baselines except Madras et. al. which matches clinician policy as it always defers.

B.1 Hyperparameters and Settings.

Method	Best parameter I	Best parameter II
SLTD	$c(0.00)$	$\tau(0.20)$
SLTD-Stationary	$c(0.10)$	$\tau(0.00)$
SLTD-One Step	$c(10.0)$	$\tau(0.00)$
SLTD ($K = 1$)	$c(0.05)$	$\tau(0.20)$
SLTD-Stationary ($K = 1$)	$c(2.00)$	$\tau(0.00)$
SLTD-One Step ($K = 1$)	$c(0.00)$	$\tau(0.00)$
Augmented-MDP	$c(0.01)$	N/A
Mozannar et. al.	$\alpha(0.20)$	N/A
Madras et. al.	$c(2.00)$	N/A

Table 9: Best parameters for all methods for Synthetic data. Corresponding values and deferral frequencies are shown in Table 1 Column I.

Method	Best parameter I	Best parameter II
SLTD	$c(0.00)$	$\tau(0.6)$
SLTD-Stationary	$c(0.05)$	$\tau(0.2)$
SLTD-One Step	$c(0.0)$	$\tau(0.6)$
SLTD ($K = 1$)	$c(0.0)$	$\tau(0.6)$
SLTD-Stationary ($K = 1$)	$c(0.1)$	$\tau(0.6)$
SLTD-One Step ($K = 1$)	$c(0.00)$	$\tau(0.6)$
Augmented-MDP	$c(0.0)$	N/A
Mozannar et. al.	$\alpha(0.2)$	N/A
Madras et. al.	$c(0.02)$	N/A

Table 10: Best parameters for all methods for Diabetes. Corresponding values and deferral frequencies are shown in Table 1 Column I.

	HIV-Case I			HIV Case-II		
Method	Self-Normalized IS (mean ± 2 s.e.)	(Best) Cost hyperparameter	Deferral Frequency (True Dynamics)	Self-Normalized IS (mean ± 2 s.e.)	(Best) Cost hyperparameter	Deferral Frequency (True Dynamics)
SLTD	16.457 ± 0.351	0.5	0.420	21.695 ± 0.217	0.5	0.346
SLTD-Stat.	17.261 ± 0.196	0.5	0.423	11.281 ± 0.397	0.5	0.289
SLTD-One Step	10.791 ± 0.331	0.2	0.487	11.227 ± 0.374	0.5	0.240
SLTD (K=1)	11.159 ± 0.517	0.5	0.449	17.864 ± 0.661	0.5	0.687
SLTD-Stat. (K=1)	8.659 ± 0.182	0.5	0.287	17.877 ± 0.157	0.5	0.889
SLTD-One Step (K=1)	6.258 ± 0.539	0.2	0.290	4.178 ± 0.131	0.5	0.518
Augmented-MDP	N/A	N/A	N/A	N/A	N/A	N/A
Mozannar et. al.	2.139 ± 0.381	0.1	0.361	3.168 ± 0.281	0.5	0.423
Madras et. al.	6.829 ± 0.116	1.0	0.687	4.173 ± 0.027	1.0	0.396
π_{tar}	3.159 ± 0.158	N/A	N/A	3.199 ± 0.212	N/A	N/A
π_0	6.158 ± 0.015	N/A	N/A	4.844 ± 0.016	N/A	N/A

Table 11: Expected Value using Self-Normalized IS for SLTD and all baselines with corresponding cost hyperparameters and deferral frequency for HIV data.

SLTD and variants. For all SLTD (original and variants), we sweep over $c \in \{0.0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$ and thresholds $\tau \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. $K = 200$ for original version and $K = 1$ for ablations. We use 5 bootstraps in each run.

Method	Best parameter I	Best parameter II
SLTD	$c(0.5)$	$\tau(0.4)$
SLTD-Stationary	$c(0.5)$	$\tau(0.0)$
SLTD-One Step	$c(0.2)$	$\tau(0.2)$
SLTD ($K = 1$)	$c(0.5)$	$\tau(0.2)$
SLTD-Stationary ($K = 1$)	$c(0.5)$	$\tau(0.0)$
SLTD-One Step ($K = 1$)	$c(0.2)$	$\tau(0.2)$
Augmented-MDP	N/A	N/A
Mozannar et. al.	$c(0.1)$	N/A
Madras et. al.	$c(1.0)$	N/A

Table 12: Best parameters for all methods for HIV-Case I. Corresponding values and deferral frequencies are shown in Table 1 Column I for Case I.

Method	Best parameter I	Best parameter II
SLTD	$c(0.5)$	$\tau(0.2)$
SLTD-Stationary	$c(0.5)$	$\tau(0.0)$
SLTD-One Step	$c(0.5)$	$\tau(0.2)$
SLTD ($K = 1$)	$c(0.5)$	$\tau(0.2)$
SLTD-Stationary ($K = 1$)	$c(0.5)$	$\tau(0.0)$
SLTD-One Step ($K = 1$)	$c(0.5)$	$\tau(0.2)$
Augmented-MDP	N/A	N/A
Mozannar et. al.	$c(0.5)$	N/A
Madras et. al.	$c(1.0)$	N/A

Table 13: Best parameters for all methods for HIV-Case II. Corresponding values and deferral frequencies are shown in Table 1 Column II for Case II.

Mozannar et. al.. This baseline has a loss penalty parameter α. We sweep over $\alpha \in \{0.0, 0.20.5, 1.0, 2.0, 5.0, 10.0\}$ and show results for the best performing α.

Madras et. al., & Augmented-MDP This baseline has a cost parameter c. We sweep over $c \in \{0.0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$ and show results for the best performing c.

All baselines were run for 5 random seeds and the average results are shown. For the deferral frequency versus value analysis, values corresponding to all choices of parameters of the respective methods are shown to demonstrate the generality of our conclusions.

B.2 Evaluation.

Value and Deferral Frequency Evaluation. To evaluate all methods we collect virtual roll-outs under the true dynamics. This is possible for Discrete Toy and Diabetes datasets. For HIV data the estimate of the dynamics are obtained using maximum likelihood estimates. We average cumulative rewards over 1000 trajectories for each method. Deferral frequency is measured as the average deferral in these trajectories.

Uncertainty Decomposition. We estimate the modeling/epistemic and irreducible/aleatoric uncertainty. This uncertainty decomposition requires the posterior estimates over the MDPs. We collect this for one sample trajectory for discrete datasets as follows. First we roll out until SLTD defers. Once we defer the first time, we simulate 10000 trajectories. Using the empirical estimates of the varianc decomposition provided in Equation 9, we can estimate all sources of uncertainty. For all baselines, the cost of deferral is constant for each time-step. Uncertainty decomposition was estimated corresponding to the best performing cost c and threshold τ.

Sample Efficiency. While our method is general, since our main focus area of applications is healthcare, we evaluate the sample efficiency of our method. Figure 9 shows the variability in Value attained as the number

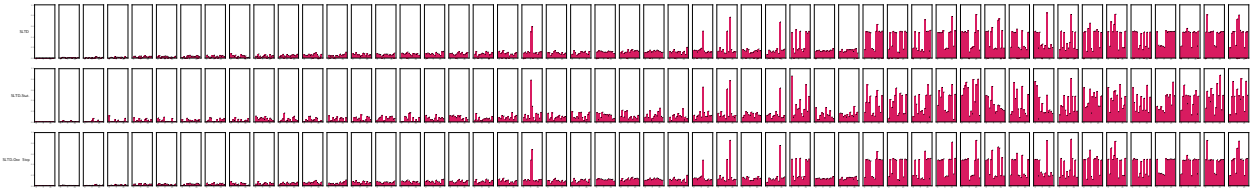


Figure 4: Learned deferral policy to Diabetes data for SLTD, SLTD-Stationary and One-Step baselines. As the dynamics shifts over time, the probability of deferring significantly increases. Minor qualitative changes across these baselines are observed suggesting Diabetes to be largely myopic (i.e., effect of interventions are observed in the near future). Modeling the non-stationarity remains crucial to obtain higher value (see Table 1).

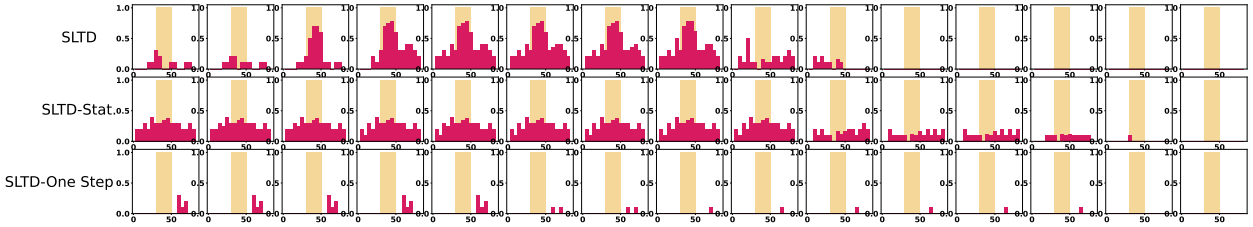


Figure 5: Learned deferral policy to HIV-I data. Shaded yellow the region of pre-emptive deferral based on clinical expertise.

of samples of off-line data available from the target environment is varied. SLTD, SLTD-Stationary, and SLTD-One Step are stable with respect to the variability in the number of samples available for training, owing to uncertainty modeling, for all datasets. Mozannar et. al. and Madras et. al. baselines are sensitive to the number of samples as they rely on an empirical risk minimization framework to learn a deferral policy. Error bars for Diabetes data are not visible due to low variability.

B.3 Computation Infrastructure

All code is implemented using Python 3.8. Discrete Toy and Diabetes experments are models were trained on a single Intel 8268 "Cascade Lake" CPUs using minimum 12GB of memory. HIV results were trained on Intel "Ice Lake" CPUs with minimum 12GB of memory. Operating system: Rocky 8. Code has also been reproduced on MacOS Monterey 12.5 (8 GB 2133 MHz LPDDR3, 2.3 GHz Dual-Core Intel Core i5 and 16 GB 3.2 GHz LPDDR4, Apple M1). Code appendix includes Anaconda package dependencies required to reproduce the results.

B.4 Additional Analysis

Evaluating learned deferral policy. To qualitatively analyze our policies, we plot the stochastic policies learned using SLTD, SLTD-One Step, SLTD Stationary in the following. Note that our policy function is $g_{\pi_{\text{tar}}}(s,t)$ is non-stationary. To visualize, we plot the probability of deferral over time. Figures 4, and 5, 6

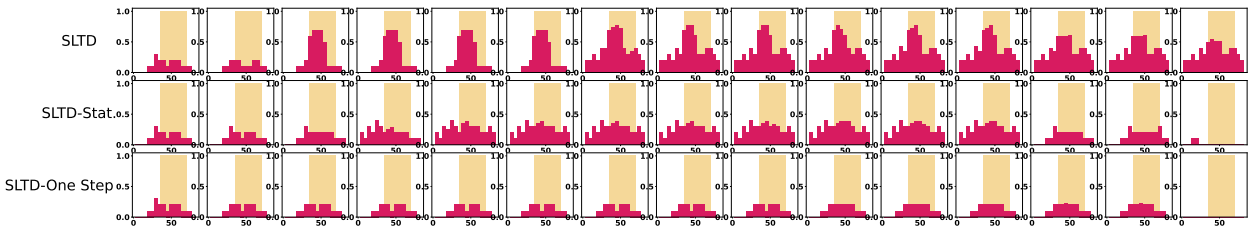
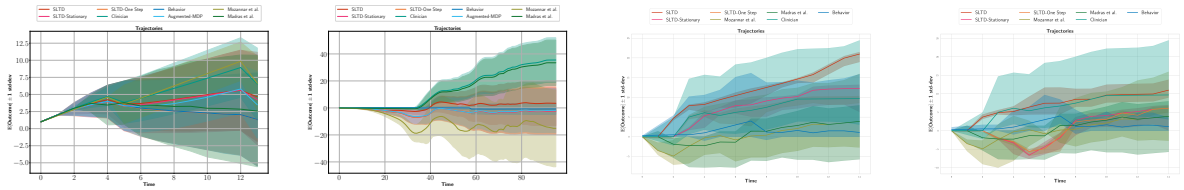


Figure 6: Learned deferral policy to HIV-II data. Shaded yellow the region of pre-emptive deferral based on clinical expertise.



Synthetic data | Diabetes | HIV-I | HIV-II

Figure 7: Sample trajectories for SLTD variants, target policy π_{tar} and clinician policy π_0. Mean improvement are comparable for discrete data but significant using SLTD for Diabetes and HIV. Relative benefits of SLTD-stationary and SLTD-one-step are less compared to modeling non-stationarity particularly for Diabetes data. Overall long-term uncertainty is comparable for all baselines for Synthetic data, Diabetes and HIV data.

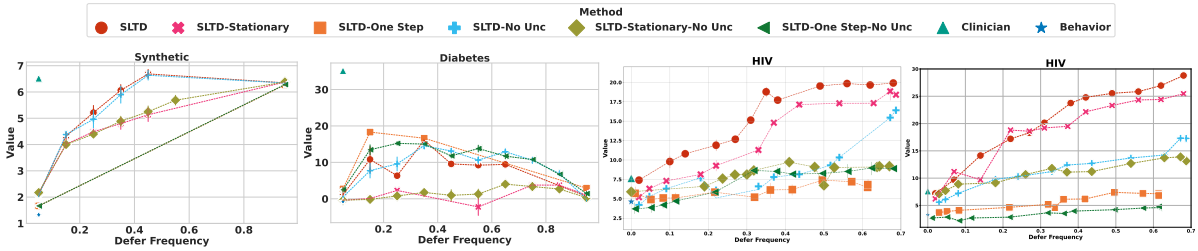


Figure 8: Trade-off of deferral frequency and expected value for SLTD, its variants, compared to ablations that do not estimate modeling uncertainty (No Unc, corresponding to $K = 1$ in Equation 3). All ablations closely follow their counterparts that leverage modeling uncertainty for Synthetic data and Diabetes. This suggests low modeling uncertainty in our environments. A higher discrepancy in performance as in HIV-Case I and Case II is indicative of higher modeling uncertainty which can occur in low data regimes. Thus accounting for this uncertainty is crucial for learning a reliable deferral policy. When we account for this uncertainty, as in the original SLTD formulations, performance is significantly better for real-world HIV data.

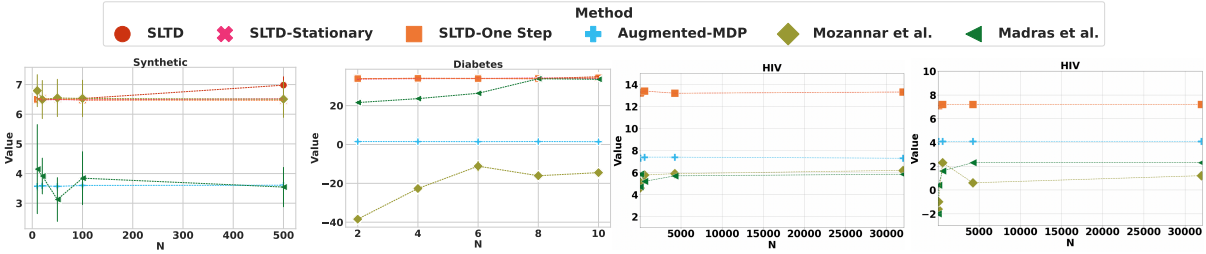


Figure 9: Sample efficiency analysis for all datasets and baselines. X-axis shows the number of episodes used for training. Y-axis shows the value obtained. For all experiments a deferral cost of 0.01 was used for consistency. SLTD, SLTD-Stationary, and SLTD-One Step are stable with respect to the variability in the number of samples available for training, owing to uncertainty modeling, for all datasets. Mozannar et. al. and Madras et. al. baselines are sensitive to the number of samples as they rely on an empirical risk minimization framework to learn a deferral policy. Error bars for Diabetes data are not visible due to low variability.

shows the learned deferral policy for SLTD, SLTD-Stationary, and SLTD-One Step for Diabetes and HIV data respectively. As can be seen in Figure 4, as the transition dynamics shift over time to that of an alternative patient, the probability of deferral increases. While SLTD and SLTD-One Step show similar learned policies qualitatively, SLTD-Stationary variant defers in different states for Diabetes data giving a sense of issues due to misspecification of the dynamics. Note that SLTD-One step does not misspecify the dynamics but defers myopically. Similarly, for HIV Case-I and II in Figures 5 and 6, the transition dynamics change over time as the virus evolves. Based on this evolution, the probability of deferral increases. In Case II, this evolution is more rapid, thus increasing the probability of deferral earlier than for Case I. Moreover, as the virus continues to mutate, the probability of deferral at subsequent steps increases for a sustained period of time in comparison to Case-I. This is typical of patients with many recombinant forms of the virus that evolve rapidly and have to be treated with more complex antiretroviral combinations.

Finally, we also notice certain differences across the two case studies for HIV. In the first case, it is evident that deferring to second line treatment in response to resistance from a first line treatment is helpful based on the results in Table 1 of the main paper. We further investigated whether deferral and points of high uncertainty correspond to certain events. In general, we observed that a higher probability of deferral and increased uncertainty at points of either virologic failure or where drug resistance has occurred. This is plausible as a change in therapy is typically required at this point to overcome such resistance. Unlike in Case I, Case II is significantly more challenging as it focuses on different viral strains that typically have a higher rate of evolution. Here, a higher probability of deferral is sustained across a longer time frame. Moreover, non-stationarity plays a significant role in this case, and the performance difference between SLTD and methods that do not account for this non-stationarity is more apparent.

Sample Trajectories. Figure 7 shows sample trajectories (with uncertainty) for all baselines and datasets to provide a sense of how the different baselines fare along with the long-term uncertainties. While no major differences between propagated uncertainties of SLTD variants is observed, the uncertainty can be higher for Mozannar et. al. especially for Diabetes data. This also applies for both cases of HIV.

Additional Analysis of Uncertainty Ablations. Table 1 shows the summary results corresponding to best performing parameters for all baselines. In addition, we also include uncertainty baselines corresponding to $K = 1$. Figure 8 shows the frequency-value trade-off for all parameter settings. For Synthetic data and Diabetes data, ablations suggest that there is no significant modeling uncertainty in our framework as the $K = 1$ or "No Unc" counterparts closely follow the performance of $K = 200$. This mainly suggests our modeling assumptions are reasonable and there is sufficient data to estimate the parameters of the dynamics resulting in low modeling uncertainty, and less variability across choice of K. This analysis can be done by collecting value estimates on true dynamics of the data for Discrete Toy and Diabetes. For HIV data, we obtain value estimates on maximum-likelihood estimates of the dynamics since true dynamics are unavailable for real-world data. Thus the analysis may be biased if the ML-estimate is biased. Nonetheless the significant difference suggests that there is indeed modeling uncertainty in the system for this data. Accounting for this uncertainty can thus have a significant impact on the long-term outcomes as it will result in potentially delayed deferrals.