Delphic Offline Reinforcement Learning under Nonidentifiable Hidden Confounding

Anonymous Author(s)
Affiliation
Address
email

Abstract

A prominent challenge of offline reinforcement learning (RL) is the issue of hidden confounding: unobserved variables may influence both the actions taken by the agent and the observed outcomes. Hidden confounding can compromise the validity of any causal conclusion drawn from data and presents a major obstacle to effective offline RL. In the present paper, we tackle the problem of hidden confounding in the nonidentifiable setting. We propose a definition of uncertainty due to hidden confounding bias, termed delphic uncertainty, which uses variation over world models compatible with the observations, and differentiate it from the well-known epistemic and aleatoric uncertainties. We derive a practical method for estimating the three types of uncertainties, and construct a pessimistic offline RL algorithm to account for them. Our method does not assume identifiability of the unobserved confounders, and attempts to reduce the amount of confounding bias. We demonstrate through extensive experiments and ablations the efficacy of our approach on a sepsis management benchmark, as well as on electronic health records. Our results suggest that nonidentifiable hidden confounding bias can be mitigated to improve offline RL solutions in practice.

1 Introduction

Large observational datasets for decision-making open the possibility of learning expert policies with minimal environment interaction. This holds promise for contexts where exploration is impractical, unethical or even impossible, such as optimising marketing, educational or clinical decisions based on relevant historical datasets [17, 62, 67]. Recent years have thus seen the emergence of offline reinforcement learning (RL) literature [41], which proposes to adapt RL methods to overcome estimation biases induced by learning from finite, fully offline data.

Aside from estimation biases, confounding variables are common in offline data [17]. The problem of hidden confounding, where outcome and decisions are both dependent on an unobserved factor, is widely overlooked in many of the concurrent offline RL methods. Nevertheless, it may induce significant errors, even for the simplest of bandit problems, and is especially aggravated in the sequential setting [7, 66, 79]. Hidden confounding exists in numerous applications. In autonomous driving, for example, the observational policy may behave according to unobserved factors (e.g. road conditions [18]), which also affect environment dynamics and rewards. Alternatively, in the medical context, unrecorded patient state information such as socio-economic factors or visual appearance may have been taken into account by the acting physician [17].

In this work, we focus on *nonidentifiable* hidden confounding in offline RL. While prior work has mostly addressed the problem in the identifiable setup [38, 43, 71, 80], we show that significant improvement in policy learning can be achieved even in the realistic nonidentifiable setting. We propose an approach to estimate uncertainty due to confounding bias and to account for the degree of

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

confoundedness while learning. In turn, this leads to improved downstream performance for offline learning algorithms.

Our main contributions are as follows. (1) To the best of our knowledge, we are the first to address *nonidentifiable* confounding bias in *deep offline RL*. (2) We achieve this by introducing a novel uncertainty quantification method from observational data, which we term delphic uncertainty. (3) We propose an offline RL algorithm that leverages this uncertainty to obtain confounding-averse policies, and (4) we demonstrate its performance on both synthetic and real-world medical data.

2 Preliminaries

We consider the contextual Markov Decision Process (MDP) [19], defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{Z}, \mathcal{A}, T, r, \rho_0, \nu, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{Z}$ is the context space, $T : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \Delta\mathcal{S}$ is the transition function, $r : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to [0, 1]$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. We assume an initial state distribution $\rho_0 : \mathcal{Z} \to \Delta\mathcal{S}$ and a context distribution ν, such that each interaction episode has a fixed context $z \sim \nu$, which may or may not be accessible to the agent, and the environment initialises at state $s_0 \sim \rho_0(\,\cdot\,|z)$. At time t, the environment is at state $s_t \in \mathcal{S}$ and an agent selects an action $a_t \in \mathcal{A}$. The agents receives a reward $r_t = r(s_t, a_t, z)$ and the environment then transitions to state $s_{t+1} \sim T(\,\cdot\,|s_t, a_t, z)$. A causal graph of the process is depicted in Figure 1.

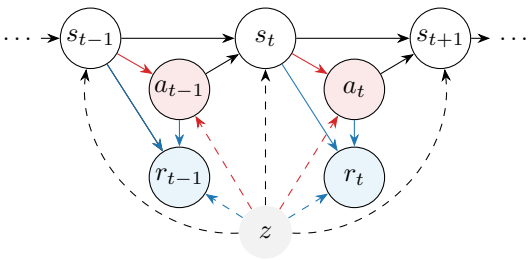


Figure 1: **Contextual MDP.** Black arrows show the transition dynamics, blues ones the reward function, and red ones the policy. Confounding arises when both behavioural policy π_b and environment returns depend on hidden context variable z (dashed lines).

We define a *context-aware* policy π as a mapping $\pi : \mathcal{S} \times \mathcal{Z} \to \Delta\mathcal{A}$, such that $\pi(a|s, z)$ is the probability of taking action a in state s and context z. Likewise, we define a *context-independent* policy as $\tilde{\pi} : \mathcal{S} \to \Delta\mathcal{A}$. We denote the set of all such policies as Π and $\tilde{\Pi}$, respectively.[1]

We assume access to a dataset of N trajectories $\mathcal{D} = \{\tau^i\}_{i=1}^N$, where the sequences $\tau^i = (s_0^i, a_0^i, r_0^i, \ldots s_H^i, a_H^i, r_H^i)$ are trajectories induced by an unknown, context-aware behavioural policy $\pi_b \in \Pi$ such that $a_t^i \sim \pi_b(\,\cdot\,|s_t^i, z^i)$. The decision-making context z^i for each trajectory is *not* included in the observational dataset. In the following, we drop index i unless explicitly needed.

Finally, we define the offline RL task with hidden confounding, which consists of finding an optimal context-independent policy $\tilde{\pi}^* \in \tilde{\Pi}$ – one which maximises the expected discounted returns. Specifically, we define the state-action value function of a policy $\tilde{\pi} \in \tilde{\Pi}$ by $Q^{\tilde{\pi}}(s, a) = \mathbb{E}_{\tilde{\pi}, z\sim\nu}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, z)|s_0 = s, a_0 = a\right]$, where $\mathbb{E}_{\tilde{\pi}}$ denotes the expectation induced by following policy $\tilde{\pi} \in \tilde{\Pi}$. We also define the value of $\tilde{\pi}$ by $V^{\tilde{\pi}}(s) = \mathbb{E}_{a\sim\tilde{\pi}}\left[Q^{\tilde{\pi}}(s, a)\right]$. An optimal policy is then defined by $\tilde{\pi}^*(\,\cdot\,|s) = \arg\max_{\tilde{\pi}\in\tilde{\Pi}}\left[V^{\tilde{\pi}}(s)\right]$.

3 Sources of Error in Offline RL

Optimising a policy from observational data is prone to various sources of error, which many RL works propose to decompose, estimate, and bound [41, 64]. First, the process is prone to statistical error in correctly estimating a value model from the observed data [28]. Inherent stochasticity in the environment (aleatoric uncertainty) can result in imprecise models, whereas finite data quantities (epistemic uncertainty) can lead to poor model approximation.

When learning from observational data, improper handling of estimation errors causes covariate shift and overestimation problems, as evident in behaviour cloning [57] and offline RL [37]. Such errors can be reduced through access to larger data quantities or online interactions with the true environment at training time. Offline RL approaches typically mitigate these errors through pessimism, penalising areas where error is expected to be large [28, 41].

[1] One can also consider history-dependent policies. Nevertheless, Markov policies sufficiently illustrate the challenges of our task, and can be easily generalised to history-dependent ones.

Another source of error, which is often overlooked in the RL literature, is structural bias. Independent of data quantity, such a bias can occur when the state-action space coverage is incomplete [68], or when the expressivity of the model class considered is inappropriate [45]. Our work considers confounding bias – a critical type of structural bias, evident in a vast number of applications [17, 18, 29, 30]. This bias can arise when the data-generating policy relies on unobserved factors that also affect downstream transitions and/or rewards [66].

Confounding Bias. Confounding bias is a critical source of error in offline RL, which is often disregarded despite many data collection environments being prone to its occurrence [29]. This source of error arises when the observational policy depends on unobserved factors which affect the chosen action and the reward or transition function. To better understand how confounding bias may affect offline RL algorithms, consider the process detailed in Section 2 and depicted in Figure 1. The offline data was generated by sampling trajectories from the behavioural policy distribution $\tau \sim P_{\pi_b}(\tau)$, which is marginalised over $\nu(z)$ and factorises as follows:

$$P_{\pi_b}(\tau) = \mathbb{E}_{z\sim\nu}\left[\rho_0(s_0|z)\prod_{t=0}^{H}\pi_b(a_t|s_t,z)\,P_r(r_t|s_t,a_t,z)\,T(s_{t+1}|s_t,a_t,z)\right], \tag{1}$$

where P_r is the probability of sampling reward r_t from $r(s_t, a_t, z)$. Any offline reinforcement learning objective can be written as an expectation over this trajectory distribution [41]. Confounding arises when one learns models on trajectories following $P_{\pi_b}(\tau)$, but estimates the value of policies π that *change* the probability of taking an action a in a given state and context (s, z) – as is necessarily the case when considering context-independent policies. Since all model terms in Equation (1) are unknown and nonidentifiable due to their dependence on z, there may exist several "worlds" that could induce the same observational distribution $P_{\pi_b}(\tau)$. This is known as the "identifiability problem" in the causal inference literature [29, 46], which has been studied extensively, providing methods for analyzing when counterfactual estimates can be obtained. Particularly, without additional assumptions about the causal structure of the environment – such as using environment interventions [43, 80] or the existence of observable back- or front-door variables [38, 71] – the context z acting as confounder is nonidentifiable and cannot be estimated. Below, we illustrate through a simple example, how two equally plausible models can correctly construct the same observational distribution, yet induce two different values for another policy.

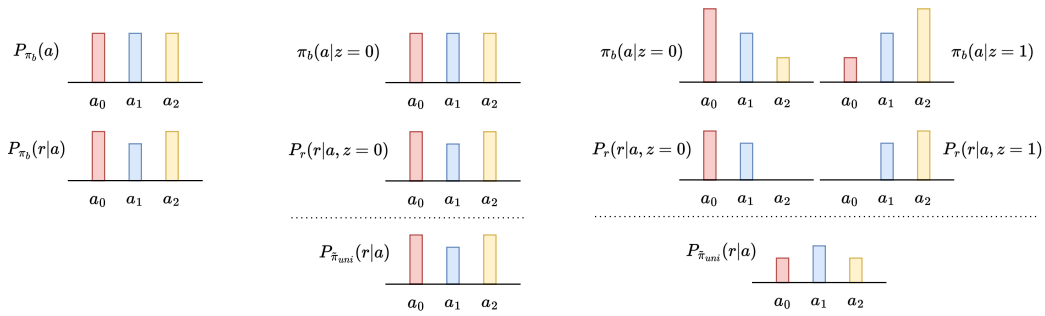
An Illustrative Example. Suppose access to the bandit data in Figure 2a, induced by an unknown context-dependent policy π_b with marginal distribution $P_{\pi_b}(a, r)$. Assume no access to the episode context z in the data. Simplifying Equation (1) to this setup, we obtain:

$$P_{\pi_b}(a, r) = \mathbb{E}_{z\sim\nu}[\pi_b(a|z)P_r(r|a, z)].$$

We can therefore change ν, π_b, and P_r to induce the same marginalised distribution P_{π_b}, with a significant difference in reward for a counterfactual policy. Indeed, in Figures 2b and 2c we show how different models that are compatible with the observational quantities can result in substantially different reward estimates for a different policy. Particularly, in World 1 (Figure 2b), we assume a deterministic singleton context, with a corresponding uniform behavioural policy, whereas in World 2 (Figure 2c) we assume two contexts with uniform distribution, and a behavioural policy which changes its distribution w.r.t. the sampled context. In both of these worlds, the observational distribution $P_{\pi_b}(a, r)$ remains the same. Nevertheless, calculating the reward of the uniform policy $\tilde{\pi}_{uni}(\cdot) = 1/|\mathcal{A}|$ results in different reward distributions. Moreover, the optimal actions in World 1 and World 2 are different.

Without explicit access to the ground-truth context or a proxy thereof (in the identifiable context), modelling an alternative policy to the privileged data-generating one will therefore be prone to spurious correlations and estimation biases.

In the next section, we propose to address the general *nonidentifiable* confounding problem in offline RL by estimating the amount of confounding error within the observational dataset and correcting for it during learning. Importantly, this source of error cannot be captured by epistemic or aleatoric uncertainty quantification methods, as discussed next.



(a) Observational data (b) World 1: $\nu(0) = 1$. (c) World 2: $\nu(0) = 0.5$.

Figure 2: **Confounding Bias Example.** World 1 and 2 are two models for the binary confounding variable that are compatible with the marginalised observational bandit data in (a), composing models for $\nu(z)$, $\pi_b(a|z)$ and $P(r|a, z)$. Under an alternative policy, such as a context-independent uniform policy $\tilde{\pi}_{uni}(\cdot) = 1/|\mathcal{A}|$, these two worlds give different values to each action.

4 Measuring Confounding Bias through Delphic Uncertainty

In this section, we formulate a method for estimating uncertainty arising from confounding bias in offline RL, which we term *delphic* uncertainty[2]. While aleatoric and epistemic uncertainty can be expressed as probability distributions over model outputs and parameters, respectively [23], delphic uncertainty is a distribution over counterfactual values. We propose a general approach to decouple aleatoric, epistemic, and delphic uncertainties, which we later leverage to overcome confounding bias in Section 5.

To introduce delphic uncertainty we first define a set of worlds compatible with the marginalised data distribution $P_{\pi_b}(\tau)$.

Definition 4.1. A compatible world for P_{π_b} is a tuple $w = (\mathcal{Z}_w, \nu_w, \rho_{0,w}, P_{r,w}, T_w, \pi_{b,w})$ which satisfies $P_{\pi_b}(\tau) = \mathbb{E}_{z\sim\nu_w}\left[\rho_{0,w}(s_0|z)\prod_{t=0}^{H}\pi_{b,w}(a_t|s_t,z)P_{r,w}(r_t|s_t,a_t,z)T_w(s_{t+1}|s_t,a_t,z)\right]$ for any trajectory $\tau = (s_0, a_0, r_0, \ldots, s_H, a_H, r_H)$. We denote by $\mathcal{W}$ the set of all compatible worlds.

We focus on uncertainty estimates of value functions. Let $w \in \mathcal{W}$ (i.e., w is some compatible world for P_{π_b}). We use θ_w to denote the parameters of a Q-value function in world w. For a fixed w and θ_w, we assume each value model Q_{θ_w} is defined by some stochastic model, e.g., a normal distribution $Q_{\theta_w}|\theta_w, w \sim \mathcal{N}\left(\mu_{\theta_w}, \sigma^2_{\theta_w}\right)$. Indeed, here σ_{θ_w} accounts for **aleatoric uncertainty**, capturing the intrinsic stochasticity of the environment [32]. Additional statistical uncertainty arises from the distribution over model parameters θ_w in the fixed world $w \in \mathcal{W}$. Starting from a prior over θ_w, evidence from the data leads to a posterior estimate over the correct model parameters $P(\theta_w|\mathcal{D})$, which captures **epistemic uncertainty** [23]. We refer the interested reader to Appendix A.1 for an overview of statistical uncertainty estimation methods.

We are now ready to define the uncertainty induced by confounding variables, which we term **delphic uncertainty**. To do this, we leverage Definition 4.1 and define delphic uncertainty by varying over compatible world models. Based on the law of total variance [73] and following on prior work separating epistemic and aleatoric uncertainty [32], we can decompose the variance in the value function estimate between the three types of uncertainties. Particularly, let w be a compatible world for P_{π_b}, and let $P_w \mapsto \Delta\mathcal{W}$ be some distribution over worlds in $\mathcal{W}$. We have the following result. Its proof, based on the law of total variance [73], is given in Appendix B.

Theorem 4.2 (Variance Decomposition)**.** *For any $\pi \in \Pi$, we have*

$$\mathit{Var}(Q^{\pi}_{\theta_w}) = \mathbb{E}_w\left[\underbrace{\mathbb{E}_{\theta_w}\left[\mathit{Var}(Q^{\pi}_{\theta_w}|\theta_w, w)|w\right]}_{\textit{aleatoric uncertainty}} + \underbrace{\mathit{Var}_{\theta_w}\left(\mathbb{E}[Q^{\pi}_{\theta_w}|\theta_w, w]|w\right)}_{\textit{epistemic uncertainty}}\right] + \underbrace{\mathit{Var}_w(\mathbb{E}_{\theta_w}[\mathbb{E}[Q^{\pi}_{\theta_w}|\theta_w, w]|w])}_{\textit{delphic uncertainty}}.$$

[2]The word "delphic" characterises quantities that are ambiguous and opaque, relating to the hidden confounding variables and their elusive effect on model predictions.

To gain further intuition of this result, consider the case of normal distributions. We can rewrite Theorem 4.2 as:

$$\mathrm{Var}(Q^{\pi}_{\theta_w}) = \mathbb{E}_w\left[\mathbb{E}_{\theta_w}[\sigma_{\theta_w}|w]^2 + \mathrm{Var}_{\theta_w}(\mu_{\theta_w}|w) + \mathrm{Var}_{\theta_w}(\sigma_{\theta_w}|w)\right] + \mathrm{Var}_w(\mathbb{E}_{\theta_w}[\mu_{\theta_w}|w]) \quad (2)$$

The first three terms, calculated by the square average of predicted standard deviations and the variance of the predicted means and standard deviations, correspond to aleatoric and epistemic uncertainties, whereas the final term, calculated by the variation over compatible world models, corresponds to delphic uncertainty. Indeed, the latter form of uncertainty cannot be diminished, even in deterministic environments and infinite data: as $|\mathcal{D}| \to \infty$ (no epistemic uncertainty), and $\sigma_{\theta_w} \to 0$ (no aleatoric uncertainty), the delphic uncertainty remains. We refer the reader to Appendix B.2 for further discussion.

5 Offline RL Under Delphic Uncertainty

In the previous section, we defined delphic uncertainty through variation over compatible world models. In this section, we propose a method to measure delphic uncertainty in practice. We then leverage our uncertainty estimate within an offline reinforcement learning framework and demonstrate its ability to mitigate confounding bias.

Following the estimation approach outlined in Theorem 4.2, delphic uncertainty can be measured through the disagreement within value functions for a given policy, under different worlds w compatible with the observational distribution.

5.1 Measuring Delphic Uncertainty

Modelling Compatible Worlds. The first practical step for evaluating delphic uncertainty is the definition of compatible world models. While one could theoretically consider all possible world models in Definition 4.1, we found that, in practice, varying over a subset of compatible models was enough to show improved offline RL efficiency.

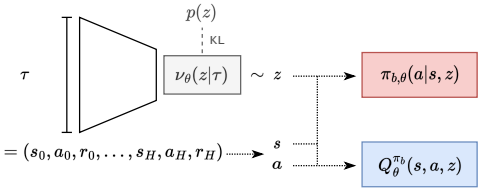


Figure 3: **Individual world model architecture** $w = (\nu_\theta, \pi_{b,\theta}, Q^{\pi_b}_\theta)$, under a prior $p(z)$ for the confounder distribution. Multiple worlds are trained and their variance in estimating Q^π is taken as delphic uncertainty.

A compatible world $w \in \mathcal{W}$ must capture key relationships from the observational data. Figure 3 depicts our proposed approach. Our model, parameterised by θ, consists of a confounder prior, a behaviour policy, and a value function estimator. During training, a trajectory $\tau \sim \mathcal{D}$ is mapped to a latent distribution $\nu_\theta(z|\tau)$, from which the policy $\pi_{b,\theta}$ and value $Q^{\pi_b}_\theta$ are estimated. Estimates are trained by draws from state-action pairs $(s, a) \sim \tau$ and a sampled z.

More specifically, we train compatible world models through variational inference, using the posterior $\nu_\theta(z|\tau)$ and prior $p(z)$. For $\tau \sim \mathcal{D}$, the model is trained by maximising the Evidence Lower BOund (ELBO, Kingma and Welling [34]):

$$\mathbb{E}_{(s,a)\sim\tau;\, z\sim\nu_\theta(z|\tau)}\left[\log Q^{\pi_b}_\theta(s,a,z) + \alpha \log \pi_{b,\theta}(a|s,z)\right] - \beta D_{KL}\left(\nu_\theta(z|\tau) \,\middle\|\, p(z)\right)$$

where $\{\alpha, \beta\}$ are hyperparameters [21] and D_{KL} is the Kullback-Leibler divergence between two distributions. Sampling from ν_θ is achieved through the reparametrisation trick [34]. Optimal parameters for $\{\nu_\theta, \pi_{b,\theta}, Q^{\pi_b}_\theta\}$ are obtained by maximising the objective over $\mathcal{D}$. Once a compatible world model is trained, the value function of a policy π can be estimated over one step using importance sampling and marginalising over z, i.e., $Q^\pi(s,a) = \mathbb{E}_{\tau\sim\mathcal{D}}\mathbb{E}_{z\sim\nu_\theta(z|\tau)}\left[\frac{\pi(a|s)}{\pi_{b,\theta}(a|s,z)} Q^{\pi_b}_\theta(s,a,z)\right]$. We discuss alternative approaches to estimating counterfactual quantities in Appendix C.

Counterfactual Variation Across Worlds. We approximate $\mathcal{W}$ by a set of W compatible worlds $\{\nu_{\theta_w}, \pi_{b,\theta_w}, Q^{\pi_b}_{\theta_w}\}_{w=1}^{W}$ on the observational training dataset $\mathcal{D}$, each trained using different priors and model architectures. Following our definition of confounding bias (Theorem 4.2), we measure delphic uncertainty through the variance in $Q^\pi_{\theta_w}(s,a)$ across worlds. That is, delphic uncertainty for policy π at state-action (s,a) is defined by $u^\pi_d(s,a) = \mathrm{Var}_w(Q^\pi_{\theta_w}(s,a))$. When no confounding exists, all models in $\mathcal{W}$ should identify similar ν, Q^{π_b} and π_b (up to epistemic uncertainty), returning

Algorithm 1: Delphic Offline Reinforcement Learning

Input: *Observational dataset $\mathcal{D}$, Offline RL algorithm.*
Learn compatible world models $\{\mathcal{Z}_w, \nu_w, \rho_{0,w}, P_{r,w}, T_w, \pi_{b,w}\}_{w \in \mathcal{W}}$ that all factorise to $P_{\pi_b}(\tau)$.
Obtain counterfactual predictions Q^π_w for each $w \in \mathcal{W}$.
Define local delphic uncertainty: $u^\pi_d(s,a) = \mathrm{Var}_w(Q^\pi_w(s,a))$.
Apply pessimism using u_d in Offline RL algorithm (see Section 5.2)

a similar value of Q^π. On the other hand, confounding with ambiguous returns would lead to different values across world models. Epistemic and aleatoric uncertainty are separately captured by implementing each world model component as an ensemble of probabilistic models. We refer the reader to Appendix C for an exhaustive overview of the training procedure.

5.2 Delphic ORL: Offline Reinforcement Learning with Delphic Uncertainty

Inspired by pessimistic approaches in offline RL [15, 28, 36, 37, 41], we propose to penalise the value of states and actions where delphic uncertainty is high, such that the learned policy is less likely to rely on spurious correlations between actions, states and rewards. This pessimistic approach, which enables the agent to account for and mitigate confounding bias when making decisions, is summarised in Algorithm 1.

In this paper, we incorporate pessimism with respect to delphic uncertainty by modifying the target Q_{target} for the Bellman update in a model-free offline RL algorithm to

$$Q'_{target}(s,a) = Q_{target}(s,a) - \lambda u^\pi_d(s,a),$$

where π is the latest learned policy, (s,a) is a tuple sampled for the update and hyperparameter λ controls the penalty strength. We apply our penalty to Conservative Q-Learning [37], but this approach could also be implemented within any model-free offline RL algorithm – which already induces pessimism with respect to epistemic uncertainty [13, 37, 41].

Note that various other methods can be adopted to drive pessimism against delphic uncertainty within existing offline RL algorithms [13, 37, 76], depending on the task at hand. The above penalty can be subtracted from the reward function in model-based methods [76]. The uncertainty measure can also be used to identify a subset of actions over which to optimise the policy, as demonstrated by Fujimoto et al. [15], or to weigh samples in the objective function, prioritising unconfounded data. We refer the reader to Appendix C for implementation details, including on the aforementioned techniques.

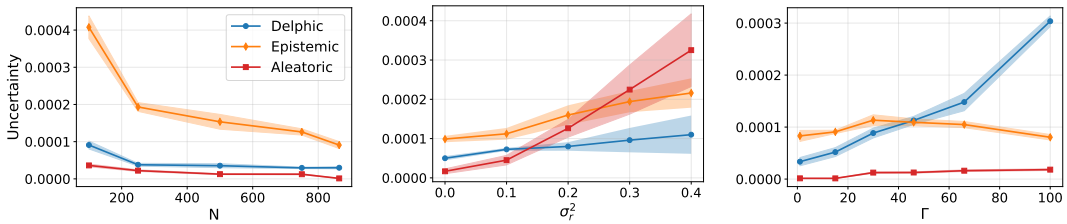
6 Experiments

In this section, we study the benefits of our proposed delphic uncertainty estimation method and its application in offline RL. We validate two principal claims: (1) Our delphic uncertainty measure captures bias due to hidden confounders. (2) Algorithm 1 leads to improved offline RL performance in both simulated and real-world confounded decision-making problems, compared to state-of-the-art but biased approaches. As baselines compatible with the discrete action spaces of environments studied here, we consider Conservative Q-Learning (CQL) [37], Batch-Constrained Q-Learning (BCQ) [15] and behaviour cloning (BC) [2]. Implementation and dataset details are provided in Appendices C and D respectively. In the following, we measure and vary confounding strength through the dependence of the behavioural policy on the hidden confounders, $\Gamma = \max_{z,z' \in \mathcal{Z}} [\pi_b(a|s,z)/\pi_b(a|s,z')]$ [56], where z also affects the transition dynamics or reward function.

6.1 Sepsis Simulation

We explore a simulation of patient evolution in the intensive care unit adapted from Oberst and Sontag [48]. The diabetic status of a patient, accessible to the near-optimal behavioural policy but absent from the observational dataset, acts as a hidden confounder z.

Uncertainty Measures. First, we study the relationship between our uncertainty estimates and the decision-making setup. In Figure 4, we find that epistemic uncertainty reduces with greater data



(a) Number of trajectories N (b) Environment stochasticity σ_r^2 (c) Confounding strength Γ

Figure 4: **Uncertainty measures as a function of data properties**, averaged over state-action pairs in the sepsis dataset. Epistemic uncertainty reduces most with more data, aleatoric uncertainty increases most with environment stochasticity (reward variance), and delphic uncertainty increases most with confounding strength.

quantities and increases out of the training set distribution, whereas aleatoric uncertainty increases with environment stochasticity, in agreement with prior work [32]. Our delphic uncertainty estimate, on the other hand, cannot be reduced with more data and increases with greater confounding. Moreover, we found that delphic uncertainty relates to meaningful regions of state-action space, as it is highest under vasopressor administration – the only treatment for which patient evolution is confounded by the hidden diabetic status. We refer the reader to Appendix E for an exhaustive overview and further experiments.

Offline RL Performance. In Figure 5, we compare environment returns obtained through offline RL, imitation learning, and our proposed approach. Our results reveal the susceptibility of offline RL to confounding bias: the presence of unobserved factors z that influence both the behaviour policy and transition dynamics leads to inaccurate value function estimates. Behaviour cloning appears to be less prone to this bias but still faces challenges in dealing with missing information in z, evidenced by the performance gap to the online policy in the unconfounded case ($\Gamma = 1$), and with the distribution shift in observed histories [50]. In contrast, our approach to penalising delphic uncertainty leads to superior performance, especially as confounding strength increases. In Appendix E, we also compare different approaches to implementing pessimism with respect to delphic uncertainty, as detailed in Section 5.2, and provide an ablation over performance as a function of pessimism hyperparameter λ.

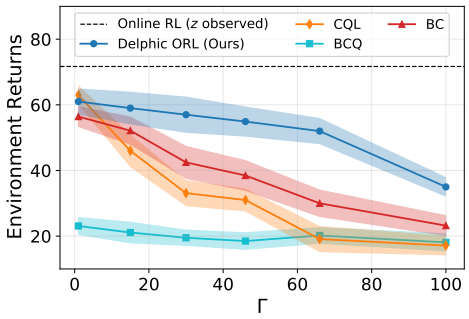


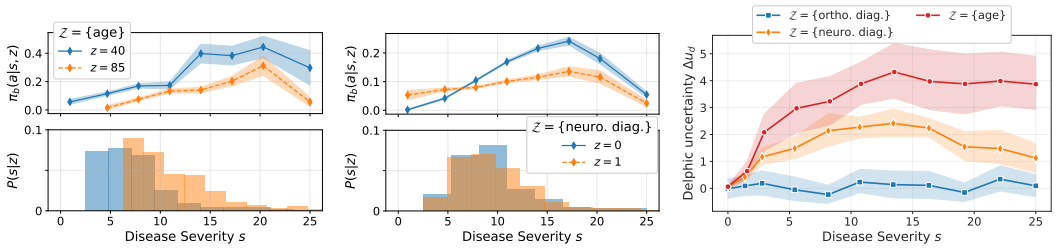
Figure 5: **Performance Results** as a function of confounding strength Γ. Normalised environment returns (mean and shaded 95% CIs) over 10 runs.

6.2 Real-World Data

We demonstrate the added value of our algorithm in optimising decision-making policies from real-world medical data. Our clinical policies are trained using a publicly available dataset of electronic health records, with over 33 thousand patient stays in intensive care and over 200 measured variables [24]. We consider the problem of optimising the treatment policy for vasopressor and fluid administration[3], and design the reward function to avoid states of circulatory failure. Significant information about patients' conditions is not available in the dataset, despite being critical to the treatment choices of attending physicians, such as socio-economic factors or medical history [75]. For ease of evaluation, we introduce additional, artificial confounders by excluding them from the observational dataset, focusing on diagnostic indicator variables, age and weight ($\Gamma \in [1, 200]$). Disease severity is measured through the SOFA score system [70].

Confounding in Medical Dataset. As the aforementioned variables affect both the probability of treatment assignment and downstream patient evolution, they act as confounders over outcome

[3]These therapeutic agents are commonly given to overcome shock in intensive care [3]. Their administration strategy has already been studied as an RL task [17, 55].



(a) {age}, $\Gamma = 48.2$ (b) {neuro. diag.}, $\Gamma = 29.1$ (c) Uncertainty increase under $\mathcal{Z}$.

Figure 6: **Delphic uncertainty as a function of** (s, z) in real-world medical data, for $a = \{\text{vasopressors}\}$. In (a, b), we note the dependence of the behavioural policy π_b (top) and/or state distribution $P(s)$ (bottom) on confounders z. In (c), delphic uncertainty increases most in confounded states and under factors with greater confounding strength, compared to orthopaedic diagnosis ($\Gamma = 3.4$).

Confounders $\mathcal{Z}$	Γ	BCQ	BC	CQL	Delphic ORL
All 14	≈200	54.6 ± 1.3	59.6 ± 0.8	59.3 ± 0.9	**62.2** ± 1.0
{age}	48.2	58.8 ± 0.8	64.7 ± 0.5	64.4 ± 0.8	**66.5** ± 0.9
{neuro. diag.}	29.1	55.0 ± 1.3	61.8 ± 0.9	59.6 ± 1.7	**65.7** ± 1.2
{gastro. diag.}	19.0	55.8 ± 0.8	60.9 ± 0.6	59.8 ± 0.6	**63.3** ± 1.1
{trauma}	16.3	56.3 ± 0.8	63.2 ± 1.1	63.5 ± 0.7	**65.7** ± 1.0
{cardio. diag.}	13.2	56.2 ± 1.0	60.6 ± 0.7	58.6 ± 0.9	**62.7** ± 1.1
{hemato. diag.}	11.6	59.6 ± 0.9	63.2 ± 0.6	63.1 ± 0.7	**65.3** ± 1.1
{weight}	8.3	60.1 ± 0.8	**64.2** ± 1.0	**65.4** ± 0.6	**66.3** ± 0.9
{sedation}	6.8	61.2 ± 0.8	**64.5** ± 0.6	**64.8** ± 0.9	**65.3** ± 1.2
{endo. diag.}	4.7	60.1 ± 1.1	63.1 ± 0.6	**65.5** ± 0.8	**65.7** ± 1.0
{resp. diag.}	4.4	61.6 ± 1.3	**64.0** ± 0.9	**65.9** ± 1.0	**64.7** ± 1.0
{ortho. diag.}	3.4	62.3 ± 0.8	**64.6** ± 0.6	**65.8** ± 0.7	**65.9** ± 1.0
{surgical status}	3.2	62.2 ± 1.1	64.3 ± 0.5	**67.4** ± 0.7	**66.8** ± 1.1
{sepsis}	2.8	60.3 ± 0.9	63.9 ± 0.6	**65.4** ± 0.7	**66.2** ± 1.0
{intoxication}	1.2	62.3 ± 0.9	63.4 ± 0.5	**65.2** ± 0.6	**66.6** ± 1.1
$\emptyset$	1	62.6 ± 0.8	**65.4** ± 0.5	**68.2** ± 0.7	**67.6** ± 1.1

Table 1: **Off-Policy Evaluation (OPE)** on the real-world medical dataset. Delphic ORL yields improvements when z strongly confounds treatment decisions (large Γ). Mean and 95% CIs over 10 runs. Best and overlapping results in bold.

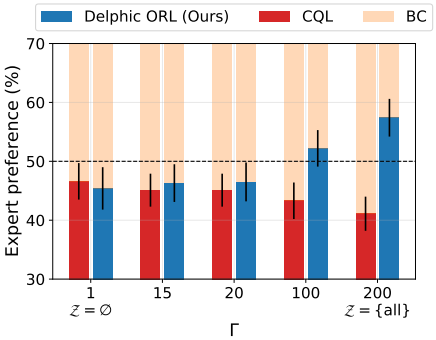


Figure 7: **Expert Clinician Evaluation** of treatment policies, supporting the conclusion that Delphic ORL improves learning in confounded settings.

models when excluded from the data. In Figure 6, we highlight how our delphic uncertainty measure captures confounded state-action pairs in concordance with the introduced confounders. Delphic uncertainty is generally highest for high disease severity, where important factors such as age or comorbidity may affect the choice of treatment intensity [1]. Indeed, delphic uncertainty increased to a greater extent under important confounders (e.g., age or patients' neurological diagnosis) than less critical factors (e.g., orthopaedic diagnosis).

Confounding-Averse Policies. We investigate the efficacy of our penalisation approach in learning policies that optimise patient outcomes in the presence of confounding bias. To evaluate the performance of these policies, we employ doubly-robust off-policy evaluation (OPE) [27, 40], which provides a confounding-independent estimate of treatment success by leveraging access to z. We refer the reader to Appendix D.2 for an exhaustive overview of this evaluation method.

Table 1 shows our approach maintains improved performance even as the confounding level increases, while offline RL methods suffer from bias and yield suboptimal policies. As an ablation, we also studied the discrepancy of our trained policy with that in the data. Particularly, we compared the actions taken by our policy and the policy in the data and found that, unlike behaviour cloning, our policy was significantly different. This suggests our learned policy was indeed able to extrapolate from the data efficiently, identifying treatment strategies that may be more robust to confounding biases. We refer the reader to Appendix E for an overview of this ablation.

Expert Clinician Evaluation. Motivated by the observed success of our method, we evaluated our algorithm using expert clinicians. Figure 7 shows the evaluation results of six human expert clinicians, who ranked pairs of different policies based on their observed patient outcomes. More specifically, the human experts were shown simulated patient trajectories and were asked to blindly compare the

expected value of actions from either our policy or the CQL policy to those of the behaviour cloning policy. The results provide additional validation for the performance improvements of our method in confounded settings. We refer the reader to Appendix D.2 for an exhaustive overview of the clinician evaluation experiment.

7 Related Work

Online-RL methods rely on environment interaction for training, limiting their applicability in many real-world domains such as healthcare [17]. This has fueled research efforts in offline methods to optimise policies through pessimism [6, 9, 28, 41, 68, 74]. Recent practical algorithmic developments in offline RL have focused on addressing statistical errors induced by epistemic and aleatoric uncertainty, in both model-based and model-free methods [13, 33, 35, 37, 76].

Structural errors such as confounding bias are also pervasive in offline RL [43]. Such biases cannot be captured by epistemic or aleatoric uncertainty quantification methods, as they do not depend on data quantity. Confounding bias cannot be reduced to the missing information problem in partially-observable environments either [20]. History-dependent policies, for example, are equally prone to this source of error: while long-term information can recover latent environment information, it exacerbates distribution shifts between behavioural and learned policies when learning from observational data [50, 63].

Several approaches have been proposed to address confounding bias in offline RL. Most make assumptions to estimate the confounding variables, including access to the environment [43, 80] or to observable back- or front-door proxy variables [38, 44, 61, 71]. This allows algorithms to apply covariate adjustment methods [54] to correct for confounding when modelling alternative policies (interventional probabilities and counterfactuals). Extensive work also discusses confounding bias in off-policy evaluation [4, 5] and bandits [8, 59, 65], but the proposed solutions remain poorly translatable to learning offline RL policies in practice, due to the aforementioned limiting assumptions.

Our work is also closely related to research on sensitivity analysis for treatment effect estimation under hidden confounding [26, 31, 49, 56]. These works propose partial identification bounds for confounded heterogeneous treatment effect estimation or bandit decision-making problems [29] by assuming a bound on the dependence of the behavioural policy on hidden confounders. In this context, Jesson et al. [26] also distinguish sources of aleatoric and epistemic uncertainty from confounding biases. Other work has proposed sensitivity analysis bounds for off-policy evaluation, formulating uncertainty sets over policy returns [46, 79]. Still, regret bounds from sensitivity analysis remain wide and often ill-adapted to high-dimensional state and action spaces or sequential decision-making problems. Our approach complements these theoretical frameworks with a practical solution to addressing confounding bias in offline RL. Finally, Saengkyongam et al. [58], Tennenholtz et al. [66] also study confounding in offline environments, but are more concerned with the complementary challenge of covariate shift – with the latter work even assuming access to the contextual information.

8 Conclusion

We proposed a practical solution to address the challenge of learning from confounded data, specifically in situations where confounders are unobserved and cannot be identified. Delphic ORL captures uncertainty by modelling world models compatible with the observational distribution, achieving improved performance across both simulated and real-world confounded offline RL tasks. Our results demonstrate that Delphic ORL can learn useful policies in cases where traditional algorithms fail due to excessive confounding. Overall, we believe research into tackling hidden confounding in offline RL will lead to more reliable and effective decision-making tools in various critical fields.

Finally, we note several limitations of our work. First, we focused our empirical evaluation on medically-motivated confounding scenarios, on the hypothesis that these should be representative of general confounded decision-making contexts. Second, the computational cost of modelling compatible worlds in Delphic ORL may be expensive for large-scale, highly confounded problems. Lastly, as with any RL algorithm, the effectiveness and safety of Delphic ORL depend on the quality and representativeness of the training data. We refer the reader to Appendix A.3 for further discussion on the limitations of our approach as well as the possible societal impact of our work.

References

[1] Élie Azoulay, Barbara Metnitz, Charles L Sprung, Jean-François Timsit, François Lemaire, Peter Bauer, Benoît Schlemmer, Rui Moreno, Philipp Metnitz, and SAPS 3 investigators. End-of-life practices in 282 intensive care units: data from the saps 3 database. *Intensive care medicine*, 35: 623–630, 2009.

[2] Michael Bain and Claude Sammut. A framework for behavioural cloning. *MACHINE INTELLIGENCE 15*, pages 103–129, 1996. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.1759`.

[3] Julia Benham-Hermetz, Mark Lambert, and Robert CM Stephens. Cardiovascular failure, inotropes and vasopressors. *British Journal of Hospital Medicine*, 73(Sup5):C74–C77, 2012.

[4] Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.

[5] Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1999–2007. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/bennett21a.html`.

[6] Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*, 2021.

[7] Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.

[8] Siyu Chen, Yitan Wang, Zhaoran Wang, and Zhuoran Yang. A unified framework of policy learning for contextual bandit with confounding bias and missing observations. *arXiv preprint arXiv:2303.11187*, 2023.

[9] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 3852–3878. PMLR, 2022.

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[11] Francesco D'Angelo, Vincent Fortuin, and Florian Wenzel. On stein variational neural network ensembles. *arXiv preprint arXiv:2106.10760*, 2021.

[12] Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982. doi: 10.1137/1.9781611970319. URL `https://epubs.siam.org/doi/abs/10.1137/1.9781611970319`.

[13] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

[14] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[15] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[17] Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

[18] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *NeurIPS*, 2019.

[19] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *CoRR*, abs/1502.02259, 2015. URL `http://arxiv.org/abs/1502.02259`.

[20] Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *2015 aaai fall symposium series*, 2015.

[21] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[23] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

[24] Stephanie L. Hyland, Martin Faltys, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26:364–373, 3 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0789-4. URL `https://www.nature.com/articles/s41591-020-0789-4`.

[25] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11637–11649. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/860b37e28ec7ba614f00f9246949561d-Paper.pdf`.

[26] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4829–4838. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/jesson21a.html`.

[27] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[28] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[29] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

[30] Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 33:22293–22304, 2020.

[31] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2281–2290. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/kallus19a.html`.

[32] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[33] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

[34] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

[35] Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR, 2021.

[36] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *NeurIPS*, 32, 2019.

[37] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[38] Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 34:14669–14680, 2021.

[39] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[40] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[41] Sergey Levine, Aviral Kumar, et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. 5 2020. URL `https://arxiv.org/abs/2005.01643v3`.

[42] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[43] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

[44] Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589*, 2022.

[45] Tyler Lu, Dale Schuurmans, and Craig Boutilier. Non-delusional q-learning and value-iteration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/5fd0245f6c9ddbdf3eff0f505975b6a7-Paper.pdf`.

[46] Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18819–18831. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/da21bae82c02d1e2b8168d57cd3fbab7-Paper.pdf`.

[47] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[48] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. *ICML*, 2019.

[49] Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. *International Conference on Machine Learning*, 2023.

[50] Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

[51] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[52] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2022.

[53] Alizée Pace, Alex Chan, and Mihaela van der Schaar. Poetree: Interpretable policy learning with adaptive decision trees. In *International Conference on Learning Representations*, 2022.

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

[55] Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo A. Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *NIPS*, 2017.

[56] Paul R Rosenbaum. Observational studies. In *Observational Studies*, pages 1–17. Springer, 2002.

[57] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `https://proceedings.mlr.press/v15/ross11a.html`.

[58] Sorawit Saengkyongam, Nikolaj Thams, Jonas Peters, and Niklas Pfister. Invariant policy learning: A causal perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[59] Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual Bandits with Latent Confounders: An NMF Approach. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 518–527. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/sen17a.html`.

[60] Takuma Seno and Michita Imai. d3rlpy: An offline deep reinforcement learning library. *Journal of Machine Learning Research*, 23(315):1–20, 2022. URL `http://jmlr.org/papers/v23/22-0017.html`.

[61] Chengchun Shi, Jin Zhu, Shen Ye, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pages 1–12, 2022.

[62] Adish Singla, Anna N Rafferty, Goran Radanovic, and Neil T Heffernan. Reinforcement learning for education: Opportunities and challenges. *arXiv preprint arXiv:2107.08828*, 2021.

[63] Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Zhiwei Steven Wu. Sequence model imitation learning with unobserved contexts. *arXiv preprint arXiv:2208.02225*, 2022.

[64] Guy Tennenholtz and Shie Mannor. Uncertainty estimation using riemannian model dynamics for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 19008–19021, 2022.

[65] Guy Tennenholtz, Uri Shalit, Shie Mannor, and Yonathan Efroni. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pages 430–439. PMLR, 2021.

[66] Guy Tennenholtz, Assaf Hallak, Gal Dalal, Shie Mannor, Gal Chechik, and Uri Shalit. On covariate shift of latent confounders in imitation and reinforcement learning. In *ICLR*, 2022. URL https://openreview.net/forum?id=w01vBAcewNX.

[67] Philip Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4740–4745, 2017.

[68] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

[69] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *1st Conference on Neural Information Processing Systems (NIPS 2017),*, 2017.

[70] J L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix), 1996.

[71] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34:21164–21175, 2021.

[72] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[73] N.A. Weiss, P.T. Holmes, and M. Hardy. *A Course in Probability*. Pearson Addison Wesley, 2006. ISBN 9780201774719. URL https://books.google.ch/books?id=Be9fJwAACAAJ.

[74] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[75] Shu Yang and Judith J. Lok. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, 28(4):1703–1723, 2018. ISSN 10170405, 19968507. URL https://www.jstor.org/stable/26511185.

[76] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

[77] Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Ratsch. Hirid-icu-benchmark — a comprehensive machine learning benchmark on high-resolution icu data. *NeurIPS*, 6 2021. URL https://physionet.org/content/hirid/1.1.1/.

[78] Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab, 2016.

[79] Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. *NeurIPS*, 32, 2019.

[80] Junzhe Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. *ICML*, page 119, 2020.

A Additional Related Work

A.1 Statistical Uncertainty Estimation

Uncertainty estimation is a crucial aspect of machine learning models, as it provides valuable insights into the reliability and confidence of model predictions and can be used to guide policy optimisation in reinforcement learning. Statistical sources of error can be estimated through aleatoric and epistemic uncertainty, which have been widely studied in the machine learning literature [23]. In this section, we review existing methodologies for capturing and quantifying these two types of uncertainty.

Aleatoric Uncertainty. Aleatoric uncertainty, also known as data uncertainty or irreducible noise, stems from the inherent variability and randomness in the observed data [23]. This form of statistical uncertainty cannot be reduced even with infinite data quantities.

The most common approach to modelling aleatoric uncertainty is to set a probability distribution over model outputs and to learn its parameters [32]. Outputs can for instance be assumed to be normally distributed with either a fixed variance (introducing a single parameter to be estimated through maximum likelihood), or a variance that depends on the input. In this latter case of heteroscedastic aleatoric uncertainty, a separate neural network branch can be trained to predict the variance [32].

Epistemic Uncertainty. Epistemic uncertainty arises from the lack of knowledge or ambiguity in the model parameters [23], which can be reduced with additional data. Capturing epistemic uncertainty is particularly important to approximate model error in out-of-distribution scenarios [28].

Bayesian neural networks (BNNs) offer a principled approach to capturing epistemic uncertainty [47]. By placing prior distributions over the model weights and using Bayesian inference, BNNs can provide posterior distributions over the weights, which represent the uncertainty in the model parameters. This uncertainty can then be propagated through the network to obtain predictive distributions that quantify epistemic uncertainty.

Bootstrap ensemble methods are another effective epistemic uncertainty estimation technique [12]. These methods rely on creating multiple subsets, or bootstrapped samples, from the original dataset by randomly sampling with replacement. Each bootstrapped sample is then used to train a separate model, resulting in an ensemble of models with slightly different parameter configurations. By aggregating the predictions from these diverse models, the epistemic uncertainty can be estimated through measures such as variance or entropy. Bootstrap ensemble methods provide a practical and scalable approach to capturing model uncertainty, particularly when Bayesian methods are computationally expensive or infeasible [39].

Monte Carlo dropout sampling [16] can also be used to estimate epistemic uncertainty by performing multiple forward passes with dropout enabled at test time. The distribution of predictions from these multiple samples gives an estimate of the predictive uncertainty. Finally, more recent efforts in epistemic uncertainty estimation include randomised priors [51], epistemic neural networks [52] and deep ensembles trained with Stein variational gradient descent [11, 42].

A.2 Nonidentifiable Confounding Bias

Comparison to Sensitivity Analysis. In the causal inference literature, sensitivity analysis studies the robustness of treatment effect estimation to hidden confounding. This framework assumes a bound on the ratio between treatment propensities between any two confounder values [56] or on the ratio between treatment propensities when accounting for and marginalising confounders [25, 31, 49].

In contrast, the main assumptions in our delphic uncertainty estimate determine which 'world models' compatible with the observational data are considered to construct uncertainty sets over a given outcome model (Conditional Average Treatment Effect or, in the sequential setting, Q-value function). In particular, we consider a set of possible $\mathcal{Z}$ and prior distributions $p(z)$, and specify a model architecture for the dependence of the behavioural policy and transition, reward or value function on z (which is then trained to fit the marginalised observational trajectory distribution).

Importantly, sensitivity analysis approaches require domain expertise to set maximum propensity-ratio parameter Γ [31], from which uncertainty sets over the modelled outcome are derived. Delphic ORL does come with its own set of hyperparameters (number of world models considered in u_d,

pessimism hyperparameter λ), which can be determined through practical, quantitative means with arguably less domain expertise.

Time-Varying Confounders. The Contextual Markov Decision Process [19] and associated problem described in Section 2 describes confounders as sampled from a context distribution $\nu(z)$ and fixed over the course of an episode. We note that this framework does not exclude the existence of time-varying confounders. Consider the Markov Decision Process with Unmeasured Confounding (MDPUC) [78], in which a new i.i.d. hidden confounder variable z_t affects the transition at each timestep t. This framework can be framed as a CMDP where the overall episode context $z = \{z_1, \ldots, z_H\}$ includes all confounder variables. Although we do not focus on this specific framework in our experimental setting, this would form an interesting avenue for further work. An important distinction between MDPUC and Partially-Observable Markov Decision Processes (POMDPs) is the assumption that confounder variables are sampled i.i.d. at each timestep. While POMDPs can therefore be viewed as the generalisation of this decision-making setup, note that confounding biases are only induced in this setup if the behavioural policy has access to some missing information about the state variable.

A.3 Broader Impact, Limitations and Future Work

Addressing hidden confounding in offline reinforcement learning has the potential to significantly impact the development and deployment of reinforcement learning systems in real-world applications. By improving the validity of causal conclusions drawn from data, Delphic ORL can improve the effectiveness and safety of RL-based decision-making in critical fields [17, 62, 67].

While our results demonstrate the efficacy of Delphic ORL in learning useful policies in the presence of confounding, it is important to acknowledge the limitations and potential unintended consequences associated with RL algorithms, especially in high-stakes applications such as healthcare. Collaboration with domain experts is crucial to ensure thorough evaluation of RL algorithms [17]. In clinical settings, predictive or recommendation models derived from Delphic ORL should not be solely relied upon, and mitigation strategies must be implemented to minimise negative consequences during deployment.

An important consideration in the application of Delphic ORL is the trade-off between confounding bias and estimation variance in Q-function estimation, as noted in other work addressing confounding bias [72]. This emphasizes the significance of large, high-quality training datasets to leverage the benefits of Delphic ORL and ensure sufficient predictive power.

As our experiments primarily focused on medically-motivated confounding scenarios, future work should investigate the applicability and generalisation of Delphic ORL to other domains. Although our framework does not in theory exclude dynamic environments where confounding factors change over time (see Appendix A.2), an empirical study of the behaviour of delphic uncertainty estimates and pessimism penalties may reveal new challenges in this context.

Finally, the question of how to best approximate the set of compatible worlds $\mathcal{W}$ in Definition 4.1 remains open. In Section 5 and Appendix C, we detail our approach which efficiently captures variability across counterfactual value-function, but further theoretical or practical work on how best to model $\mathcal{W}$ would likely improve the calibration of delphic uncertainty estimates. Better approximation algorithms may also improve the efficiency, scalability, and modelling power of our method for very high-dimensional, highly confounded problems – although our real-world data analysis forms a promising first proof-of-concept.

B Theoretical Details

B.1 Proof of Theorem 4.2

We start by considering the decomposition of variance in Q^π_θ caused by random variable θ. In the following, we drop superscript π for clarity.

First, we decompose $\text{Var}(Q_\theta \mid \theta)$:

$$\begin{aligned}\text{Var}(Q_\theta \mid \theta) &= \mathbb{E}[Q_\theta^2|\theta] - \mathbb{E}[Q_\theta|\theta]^2 \\ \mathbb{E}_\theta[\text{Var}(Q_\theta \mid \theta)] &= \mathbb{E}_\theta[\mathbb{E}[Q_\theta^2|\theta]] - \mathbb{E}_\theta[\mathbb{E}[Q_\theta|\theta]^2] \\ &= \mathbb{E}[Q_\theta^2] - \mathbb{E}_\theta[\mathbb{E}[Q_\theta|\theta]^2] \end{aligned} \tag{3}$$

where the last line results from the law of iterated expectations: $\mathbb{E}_B[\mathbb{E}[A|B]] = \mathbb{E}[A]$ for two random variables A, B.

Next, we study $\text{Var}(\mathbb{E}[Q_\theta \mid \theta])$:

$$\begin{aligned}\text{Var}_\theta(\mathbb{E}[Q_\theta \mid \theta]) &= \mathbb{E}_\theta[\mathbb{E}[Q_\theta \mid \theta]^2] - \mathbb{E}_\theta[\mathbb{E}[Q_\theta \mid \theta]]^2 \\ &= \mathbb{E}_\theta[\mathbb{E}[Q_\theta \mid \theta]^2] - \mathbb{E}[Q_\theta]^2 \end{aligned} \tag{4}$$

again using iterated expectations.

Summing equations 3 and 4, we obtain:

$$\begin{aligned}\mathbb{E}_\theta[\text{Var}(Q_\theta \mid \theta)] + \text{Var}_\theta(\mathbb{E}[Q_\theta \mid \theta]) &= \mathbb{E}[Q_\theta^2] - \mathbb{E}[Q_\theta]^2 \\ &= \text{Var}(Q_\theta) \end{aligned} \tag{5}$$

This result is known as the law of total variance [73], which can be interpreted as a decomposition of epistemic and aleatoric uncertainty [32].

We can rewrite the above result within a given world model w, denoting θ as θ_w. Now conditioning on the world model w, we have:

$$\text{Var}(Q_{\theta_w} \mid w) = \mathbb{E}_{\theta_w}[\text{Var}(Q_{\theta_w} \mid \theta_w, w)|w] + \text{Var}_{\theta_w}(\mathbb{E}[Q_{\theta_w} \mid \theta_w, w]|w) \tag{6}$$

We also write equation 5 such that the conditioning random variable is now w, which induces variation in Q_{θ_w} if we consider a counterfactual trajectory distribution. Combined with Equation (6), we obtain:

$$\begin{aligned}\text{Var}(Q_{\theta_w}) &= \mathbb{E}_w[\text{Var}(Q_{\theta_w}|w)] + \text{Var}_w\left(\mathbb{E}[Q_{\theta_w} \mid w]\right) \\ &= \mathbb{E}_w\Big[\mathbb{E}_{\theta_w}[\text{Var}(Q_{\theta_w} \mid \theta_w, w)|w] + \text{Var}_{\theta_w}(\mathbb{E}[Q_{\theta_w} \mid \theta_w, w]|w)\Big] \\ &\qquad + \text{Var}_w\Big(\mathbb{E}_{\theta_w}\left[\mathbb{E}[Q_{\theta_w} \mid \theta_w, w]|w\right]\Big) \end{aligned} \tag{7}$$

using iterated expectations. This concludes the proof of Theorem 4.2.

Note that if we assume Q_{θ_w} has a Gaussian distribution for fixed $\{\theta_w, w\}$, parameterised as $\mathcal{N}(\mu_{\theta_w}, \sigma^2_{\theta_w})$, we have $\text{Var}(Q_{\theta_w} \mid \theta_w, w) = \sigma^2_{\theta_w}$ and $\mathbb{E}[Q_{\theta_w} \mid \theta_w, w] = \mu_{\theta_w}$. We obtain results in Equation (2) by expanding the first term in the variance decomposition, $\mathbb{E}_{\theta_w}[\sigma^2_{\theta_w}|w]$, as follows:

$$\begin{aligned}\mathbb{E}_{\theta_w}[\sigma^2_{\theta_w}|w] &= \mathbb{E}_{\theta_w}[\sigma^2_{\theta_w}|w] - \mathbb{E}_{\theta_w}[\sigma_{\theta_w}|w]^2 + \mathbb{E}_{\theta_w}[\sigma_{\theta_w}|w]^2 \\ &= \text{Var}_{\theta_w}(\sigma_{\theta_w}|w) + \mathbb{E}_{\theta_w}[\sigma_{\theta_w}|w]^2. \end{aligned}$$

B.2 Asymptotic Interpretation of Theorem 4.2

We consider three extreme cases of Theorem 4.2 to clarify its decomposition. First, we consider the limit of infinite-data with no confounding (e.g., no dependence on z). In this case, θ_w and w converge to a single ground-truth. Any remaining statistical error will come from the intrinsic environment stochasticity or the behavioural policy, and therefore has an aleatoric nature. Indeed, only the first term in Theorem 4.2 would remain.

Next, consider the setting in which the value is a deterministic mapping of states, with only one compatible world model. Learning from finite data quantities leads to statistical error in optimising the parameters θ_w, and is known as epistemic uncertainty. Indeed, deterministic environments with only one compatible world model will reduce Theorem 4.2 to the second term.

Finally, we consider the case of infinite data in a deterministic setting. In this case, multiple compatible world models may exist which induce the same observational distribution (as demonstrated in

Section 3). The source of error remaining is delphic uncertainty, and arises if multiple models assign high likelihood to the observational data, but return different estimates of the value. In this paper we propose to estimate this final form of uncertainty by learning an ensemble of compatible world models, in a similar fashion to the bootstrap method for quantifying epistemic uncertainty.

C Implementation Details

C.1 Statistical & Delphic Sources of Uncertainty

World Model Training. We implement world models as variational models for estimating the confounder distribution, jointly with a model for the behaviour policy π_b and for the action-value function Q^{π_b}, both dependent on a z sampled from the posterior. As the environments we consider have discrete action spaces, we learn the behaviour policy by minimising its cross-entropy on the training data, as in behaviour cloning. Training is carried out for 50 epochs or until loss on the validation subset (10% of training data) increases for more than 5 consecutive epochs. Within a world model w, hyperparameters $\{\alpha, \beta\}$ can be tuned based on prediction performance on the validation set.

Model Q^{π_b} corresponds to an on-policy action-value function approximation. We compute targets through Monte Carlo updates (for the sepsis environment with sparse episodic rewards) or Temporal Difference learning (for the real-world ICU dataset) based on samples from the observational training data with a discount factor of $\gamma = 0.99$. The Q-function is trained as a classifier over 200 quantiles.

Between world models w, the confounder space dimensionality is randomly varied over $|\mathcal{Z}| = \{1, 2, 4, 8, 16\}$, and the prior for $p(z) = \mathcal{N}(z; 0, \Sigma^2)$ is randomly varied through the variance for each z-dimension, $\Sigma^2_{ii} = \{1.0, 0.1, 0.01\}$. For the sepsis simulation, the encoder architecture for the confounder distribution $\nu(z|\tau))$ consists of a multi-layer perceptron with hidden layer dimensions $(128, 64, 32)$ and ReLU activation before the final layer mapping to dimension $|\mathcal{Z}|$. For the real dataset, the encoder architecture is implemented as a transformer [69] with 2 layers, 4 heads, and embedding dimension 32, considering a maximum history length of 10 tokens. The behavioural policy $\pi_b(a|s, z)$ and action-value function $Q^{\pi_b}(s, a, z)$ are both implemented as multilayer perceptrons with hidden layer dimensions $(32, 64, 128)$ and ReLU activation.

Uncertainty Estimates. Additional inductive biases can be incorporated to capture epistemic and aleatoric uncertainty within a single world model w, as these relate to statistical sources of uncertainty. Following prior work [32, 76], we capture aleatoric uncertainty by modelling a normal probability distribution over outputs (π_b, Q^{π_b}). We then measure epistemic uncertainty within each world model w by training on different data bootstraps, returning an ensemble of parameters $\{\theta^1_w, \theta^2_w, \ldots\}$ for each w.

Recalling Equation (2), the delphic uncertainty term $\text{Var}_w(\mathbb{E}_\theta[\mu_{\theta_w}])$ is estimated by measuring the variance between predictions μ_{θ_w} (averaged over model parameters θ_w), across across multiple generative models w. Epistemic uncertainty can be estimated as the variance of outputs over different model parameters θ_w, averaged across worlds $w \in \mathcal{Z}$. Finally, aleatoric uncertainty is measured through the fitted probability distribution over model outputs $Q^{\pi_b}_{\theta_w}$, averaged over all θ_w in a given world, and over all worlds $w \in \mathcal{W}$.

The number of world models W was varied between 5 and 20 for both datasets and was chosen as the smallest number converging to an average delphic uncertainty comparable to the largest W. An ablation of delphic uncertainty as a function of the number of world models is given in Appendix E. This resulted in 10 and 15 world models for the sepsis and real-world datasets respectively. Finally, each world model was trained over 5 different data bootstraps to estimate epistemic uncertainty. Overall, compared to sensitivity analysis where parameter Γ needs to be fixed through domain expertise [49], we found delphic uncertainty to be less dependent on expert input in determining model parameters.

Counterfactual Estimates. Note that while our approach changes the policy term in P_{π_b} to obtain counterfactual estimates, other factors in the world model (e.g. ν_w, $Q^{\pi_b}_w$) could be varied to obtain general counterfactual predictions in this world model. As an example, we also found promising results by measuring delphic uncertainty through variation across w over the following counterfactual

Algorithm 2: Delphic Offline Reinforcement Learning: Bellman Penalty in Offline Q-Learning Algorithm.

1 **Input:** *Observational dataset $\mathcal{D}$, Model-free Offline RL algorithm (e.g. CQL [37]), Penalty hyperparameter λ.*
2 Learn a set of compatible world models $\{\mathcal{Z}_w, \nu_w, \rho_{0,w}, P_{r,w}, T_w, \pi_{b,w}\}_{w\in\mathcal{W}}$ that all factorise to $P_{\pi_b}(\tau)$.
3 Obtain counterfactual predictions Q^π_w for each $w \in \mathcal{W}$.
4 Define local delphic uncertainty: $u^\pi_d(s,a) = \mathrm{Var}_w(Q^\pi_w(s,a))$.
5 Initialise Q-function parameters ϕ.
6 **for** *each iteration* **do**
7 Sample $(s,a,r,s') \sim \mathcal{D}$.
8 Compute penalised Bellman target: $Q'_{target} = r + \gamma \max_{a'\in\mathcal{A}} Q_\phi(s',a') - \lambda u^\pi_d(s,a)$, where $\pi(a|s) = \mathrm{argmax}_a Q_\phi(s,a)$.
9 Perform gradient descent w.r.t. ϕ on $\left[Q_\phi(s,a) - Q'_{target}(s,a)\right]^2 + \mathcal{R}_{offline}(\phi)$, where regularisation term $\mathcal{R}_{offline}$ depends on the choice of offline learning algorithm.
10 **end**

quantity: $\mathbb{E}_{(s,a)\in\mathcal{D}}\mathbb{E}_{z\sim p_w(z)}\mathbb{E}_{\theta_w}[Q^{\pi_b}_{\theta_w}(s,a,z)]$, where z is sampled from the model prior $p_w(z)$ instead of the learned posterior $\nu_{\theta_w}(z|\tau)$. In this case, the resulting delphic uncertainty estimate, capturing variation over the counterfactual quantity across world models, becomes independent of a given policy – and dependent on the new quantity introduced (in the previous example, on prior $p_w(z)$).

C.2 Delphic Offline Reinforcement Learning

We detail our learning procedure in Algorithm 2. As our base offline RL algorithm is CQL [37], our regularisation term $\mathcal{R}_{offline}(\phi)$ is the CQL penalty: $\mathcal{R}_{offline}(\phi) = \alpha\left[\log\sum_{\tilde{a}\in\mathcal{A}}\exp Q_\phi(s,\tilde{a}) - Q_\phi(s,a)\right]$. We base our algorithm on an existing implementation for CQL [60], which includes additional training details for stability, such as target networks, double Q-networks and delayed updates [14]. For architecture details, see the baseline implementation of CQL in Appendix C.3. As for all baseline algorithms, we train for 100 epochs, using 500 (sepsis dataset) or 10^4 (ICU dataset) timesteps per epoch. In practice, the policy π considered for uncertainty estimation and the target network are updated every 8000 timesteps, to improve stability in training.

Note that an actor-critic variant of Algorithm 2 is also feasible, setting π in u^π_d to be the actor policy, as well as other offline learning paradigms in $\mathcal{R}_{offline}(\phi)$, such as BC regularisation [13].

Alternative Forms of Pessimism. Following the discussion on alternative forms of pessimism in Section 5.2, we propose practical alternatives to the Delphic ORL penalty in Line 8 of Algorithm 2, which substracts a factor of u_d from the Q-function Bellman target based. In the following, note that u_d can also be independent of π if varying over different factors in P_{π_b} as detailed above.

- **Delphic ORL via Uncertainty Threshold:** One approach, inspired by Batch Constrained Q-Learning [15], is to constrain value function updates to only consider actions falling below a certainty uncertainty threshold. For a tuple (s,a,r,s'), the Q-function Bellman target can be computed as: $Q'_{target} = r + \gamma \max_{a':u^\pi_d(s',a')<\lambda} Q_\phi(s',a')$, where λ is a threshold controlling the maximum delphic uncertainty accepted for a given action choice.
- **Model-Based Delphic ORL:** In model-based methods, a penalty proportional to the uncertainty $u_d(s,a)$ can be substracted from the reward function $r(s,a)$, as in Yu et al. [76]. The effective reward function becomes: $\tilde{r}(s,a) = r(s,a) - \lambda u_d(s,a)$.
- **Delphic ORL via Weighting:** The uncertainty measure can also be used to weight samples in the objective function, prioritising unconfounded states and actions during training:

$$\mathbb{E}_{(s,a,r)\sim\mathcal{D}}\left[\frac{\lambda}{u_d(s,a)}\mathcal{L}(s,a,r)\right]$$

where $\mathcal{L}$ can be the Q-function Bellman update or the supervised learning objective for behaviour cloning.

We compare the performance of different implementations of pessimism on the simulated sepsis environment in Appendix E.

Hyperparameter Tuning. There is no natural validation criterion in Offline RL, and the best approach to choose hyperparameters in this context remains an open question [41]. In practice, we run our algorithm for 4 different values of $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and choose the final policy giving the best off-policy evaluation performance on the validation set (using the Fitted Q-Evaluation implementation available in the codebase, Le et al. [40]). As noted in related works, expert input may be useful at this stage to also determine how strong a penalty again potential hidden confounding would be desirable or how much confounding could be expected [56]. Other hyperparameters specific to offline RL algorithms are tuned in the same way and are given in the following section.

C.3 Baseline Methods & Training Details

All reinforcement learning algorithms and baselines are implemented based on the open access `d3rlpy` library [60]. The discount factor used is $\gamma = 0.99$, and state and actions are normalised to mean 0 and variance 1 [13] for all algorithms. Training is carried out on NVIDIA RTX2080Ti GPUs on our local cluster, using the Adam optimiser with default learning rate and a batch size of 32. Models are trained for 100 epochs with 500 (sepsis dataset) or 10^4 (ICU dataset) timesteps per epoch. Model-specific hyperparameters are tuned as in Delphic ORL.

Behaviour Cloning (BC). Behaviour cloning [57] is a supervised learning model of the behaviour policy, mapping states to actions observed in the dataset. After considering the following architectures: multi-layer perceptron (MLP), Long Short Term Memory (LSTM) network [22], Gated Recurrent Unit (GRU) [10] and Transformer [69], GRU was found to give the best validation performance on both the simulated and real datasets. Implementation details for the GRU BC models include two hidden layers of dimension $(64, 32)$ and ReLU activation. The last layer is passed through a softmax layer to produce action probability outputs, and the model is trained by minimising action cross-entropy over the observational dataset, with L2 regularisation of weight 0.01.

Conservative Q-Learning (CQL). Discrete CQL [36] is implemented with a penalty hyperparameter α of 1.0 (sepsis environment) and 0.5 (ICU dataset), tuned over the following values: $\{0.1, 0.5, 1.0, 2.0, 5.0\}$. The Q-function is implemented as a distributional model with a standard MLP architecture (two linear layers with 256 hidden units) and 200 quantile regression outputs.

Batch Constrained Q-Learning (BCQ). Discrete BCQ [15] is implemented with a threshold for action flexibility set to 0.5 for both environments, tuned over the following values: $\{0.1, 0.3, 0.5, 1.0, 2.0, 5.0\}$. The Q-function is implemented as a distributional model with a standard MLP architecture (two linear layers with 256 hidden units) and 200 quantile regression outputs.

D Experimental Details

D.1 Decision-Making Environments

Sepsis Environment. Introduced by Oberst and Sontag [48], this environment simulates the trajectory of patients in the intensive care. Based on the authors' publicly available code[4], our state space $\mathcal{S}$ consists of 4-dimensional observation vectors (measures for heart rate, systolic blood pressure, oxgenation and blood glucose levels) which we normalise to mean and variance $(0, 1)$. The discrete action space $\mathcal{A}$ comprises the combination of three binary treatments (antibiotic, vasopressor or ventilation administration) for a total dimension of 8. An unobserved binary variable z encodes the diabetic status of patients, with 20% of trajectories having a positive status. The agent obtains a reward of $+1$ if the patient reaches a healthy state (and is thus ready for discharge) and a negative reward of -1 if the patient reaches a death state.

The observational dataset $\mathcal{D}$ is generated by rolling out the optimal (diabetes-aware) policy in the environment for 10,000 environment interaction steps, taking a random action with probability

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

$\epsilon = 0.1$ to ensure sufficient state-action coverage for offline learning. The maximum episode length is set to 20 timesteps. The resulting dataset has a confounding strength of $\Gamma = 100$.

Environment stochasticity can be varied by changing the variance around the originally deterministic reward obtained at the end of a trajectory, between $\sigma_r^2 = 0$ as in the original environment and $\sigma_r^2 = 0.4$. Datasets of varying confounding strength $\Gamma \in [1, 100]$ are obtained by setting the behaviour policy for $z = 1$ as a weighted average of the policies for different z values: $(1-p)\pi_b(z=0) + p\pi_b(z=1)$, where p depends on Γ and ϵ. Environment transition and reward functions and their dependence on z are kept fixed. Finally, we vary the dimension of the confounder space $\mathcal{Z}$ by introducing more binary indicators with the same effect on the transition dynamics as the diabetes indicator.

Electronic Health Records Dataset. Our real-world data experiment is based on the publicly available HiRID dataset [24]. This dataset counts over 33 thousand patient admissions at an intensive care unit in Bern University Hospital, Switzerland [24] and can be pre-processed using open access code from the HiRID benchmark [77]. Patient stays were downsampled to hourly measurements and truncated to a maximum length of 20 hours and default training, validation and test sets were used.

We consider the task of optimising fluid and vasopressor administration ($\mathcal{A}$ is the combination of two binary choices). The reward function is designed to penalise circulatory failure events ($r = -1$ for all timepoints in the duration of the event) and to reward timepoints where the patient is not in such a critical state ($r = 1$, and $r = 2$ in the timepoint following recovery from circulatory failure). Circulatory failure events for each patient are labelled following internationally accepted criteria [77]. This short-term reward function is dense, unlike previous RL work on optimising intravenous fluid and vasopressor administration [55], making off-policy evaluation more reliable [17].

The state space $\mathcal{S}$ consists of all variables in the electronic health records which are not considered treatment for the organ system considered, based on the variable categorisation released with the dataset [24]. This results in a state space dimensionality of 203. The list of variables excluded for each task in given in Table 2. At each timepoint within a patient stay, we also compute the Sequential Organ Failure Assessment (SOFA) score [70] which is used to quantify the severity of a patient's illness in the intensive care unit. A higher score indicates greater severity of illness.

Selected confounders are obtained by excluding some state dimensions from the observational dataset (up to $|\mathcal{Z}| = 14$). These variables do not constitute the *entire* confounder space, as much exogenous, unrecorded information affects patient evolution and is taken into account in medical treatment decisions [75]. We ignore this in our analysis as we cannot evaluate with respect to this missing information, but we note that this is precisely the motivation behind our work.

The confounding strength Γ for each confounding space $\mathcal{Z}$ considered was estimated as follows. Each point in the training dataset was binned into a (s, a, z) category, depending on its discrete action and context values (a, z) and on its SOFA score as a summary variable for s. We discretise the SOFA score into 5 quantiles. Finally, we compute the mean policy value for each (s, a, z) bin through $\pi_b(a|s, z) = P(a, s, z)/P(s, z)$, and we take Γ as the ratio $\max_{z,z'}[\pi_b(a|s, z)/\pi_b(a|s, z')]$.

D.2 Analysis Details

In this section, we provide additional details pertaining to the analysis of our experimental results. All results reported in this work include 95% confidence intervals around the mean, computed over ten training runs unless otherwise stated. Environment returns and off-policy evaluation results are normalised on a scale of 0 to 100. Figure 4 was obtained by varying the dimension on the x-axis, while keeping the other variables fixed to $N = 864$ trajectories, confounding strength $\Gamma = 15$ and reward function variance $\sigma_r^2 = 0.0$. To generate Figure 6, patients with the relevant confounder (z) value were binned by disease severity and the probability of vasopressor prescription (top) and the overall density (bottom) in each group were computed. Figure 6c was then obtained by computing the relative increase in delphic uncertainty when including the relevant z-dimension to the hidden context space $\mathcal{Z}$ (in other words, removing this dimension from the visible state space).

Off-Policy Evaluation (OPE). Doubly robust methods trade off bias of an approximate reward model and of weighted methods with the high variance of importance sampling approaches [27]. Assuming z is accessible for each trajectory at *evaluation* time to overcome confounding, doubly-

Table 2: **Offline reinforcement learning task on real-world medical dataset.**

Task	Circulatory treatment	
Action space $\mathcal{A} = \{0, 1\}^2$	Fluids	Vasopressors
Organ failure avoided by R	Circulatory failure	
State space $\mathcal{S}$ (selected variables, $\lvert\mathcal{S}\rvert = 204$)	Heart rate Body temperature Blood pressure Cardiac output Oxygen saturation Lactate PaO2 Serum sodium Haemoglobin Other lab values Antibiotics Diuretics Cerebrospinal fluid drain	Respiratory rate Urinary output GCS score Central venous pressure Base excess Arterial pH Creatinine Serum potassium Glucose Ventilator settings Steroids Insulin Anticoagulants ...
Other treatment variables excluded	Blood product infusions Cristalloid infusion Colloid infusion	Vasodilators Antiarrhymic agents Antihypertensive agents
Confounder variables z	Age Weight Gastrointestinal diagnosis Neurological diagnosis Hematology diagnosis Sedation Emergency status	Cardiovascular diagnosis Pulmonary diagnosis Orthopaedic diagnosis Metabolic/endocrine diagnosis Trauma diagnosis Intoxication Surgical status

robust off-policy evaluation estimates the value of policy $\tilde{\pi}$ as follows:

$$V_{DR}(\tilde{\pi}) = \mathbb{E}_{(s,a,r,z)\in\mathcal{D}}\left[\frac{\tilde{\pi}(a|s)}{\hat{\pi}_b(a|s,z)}\{r - Q(s,a,z)\} + Q(s,\tilde{\pi}(s),z)\right], \tag{8}$$

where $\hat{\pi}_b$ is a model for the behavioural policy and Q for expected returns under $\tilde{\pi}$, learned on the dataset with observable z.

Fitted Q-Evaluation is an established value estimation method [40]. The algorithm iteratively applies the Bellman equation to compute bootstrapping targets for Q-function updates: $Q_{k+1} \leftarrow \arg\min_Q \mathbb{E}_{(s,a,r,z)\in\mathcal{D}}\left[\{r - Q(s,a,z) + \gamma Q_k(s',\tilde{\pi}(s'),z)\}^2\right]$ which can be solved as a supervised learning problem. This results in a learned Q-value for the evaluated policy $Q^{\tilde{\pi}}(s,a,z)$ which can be used in the weighted doubly-robust estimate in Equation (8) to provide return estimates in Table 1.

Both the Q-function and the behaviour policy in Equation (8) are parametrised as a fully-connected neural network dimension with 3 layers of hidden dimension (64, 32, 16) and ReLU activation. The former is trained by minimising the mean squared error with the Q-function update above, the latter by minimising the cross-entropy with respect to action choices in $\mathcal{D}$.

Human Policy Evaluation. Off-policy evaluation has limitations, being itself prone to its own set of statistical errors and data-related concerns [17]. We aim to confirm conclusions drawn over OPE returns through a human expert evaluation of treatment policies.

Synthetic patient trajectories are first generated by randomly sampling from the ICU dataset along each state dimension, with varying amounts of contextual information as detailed in Table 3. Action choices at the end of the trajectories are computed for the Delphic ORL, CQL and BC policies, trained on the observational dataset with the same degree of confounding. Trajectories are selected if they induced a disagreement between these methods, to shed light on potential improvements or harmful

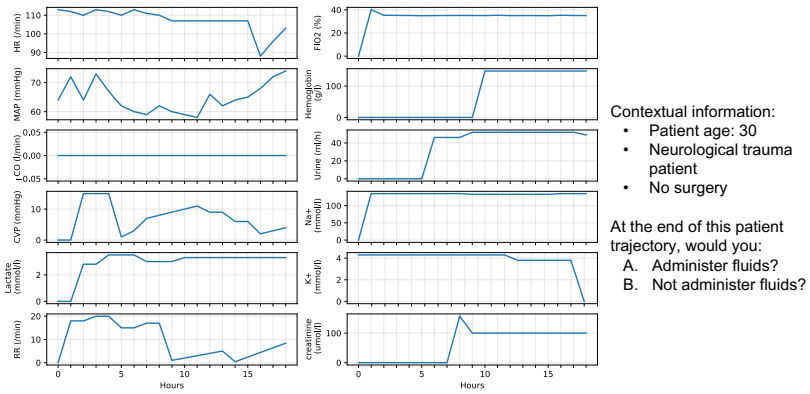


Figure 8: **Illustration of action ranking by medical experts.** Synthetic patient trajectories and a varying degree of contextual information (varying $|\mathcal{Z}|$ and Γ) are given to clinicians, who must rank the treatment options in terms of expected patient outcomes.

behaviour learned by the offline RL models. Trajectories are then simplified into 12 critical variables (as shown in Figure 8), and shown to physicians, who are asked to rank two treatment options in terms of expected patient outcomes. Unknown to the physicians, and in a random order, one of the options was predicted by the Delphic ORL or CQL policy, and the other by the BC baseline. Overall, we consulted six clinicians with different degrees of expertise in intensive care (from junior assistant doctors to department heads) from Switzerland and the United Kingdom, collecting their treatment preferences over 45 such trajectories.

Table 3: **Data settings considered during expert clinician evaluation.** Physicians are asked to rank action choices based on only state information ($\Gamma \approx 200$), or with varying amounts of observed contextual information ($\{$All 14$\}$ refers to all possible $\mathcal{Z}$ variables outlined in Table 2).

Γ	1	15	20	100	200
$\lvert\mathcal{Z}\rvert$	0	10	11	13	14
Observed	$\{$All 14$\}$	$\{$Age, Neuro. diag., Trauma diag., Surgery$\}$	$\{$Age, Neuro. diag. Surgery$\}$	$\{$Age$\}$	$\emptyset$

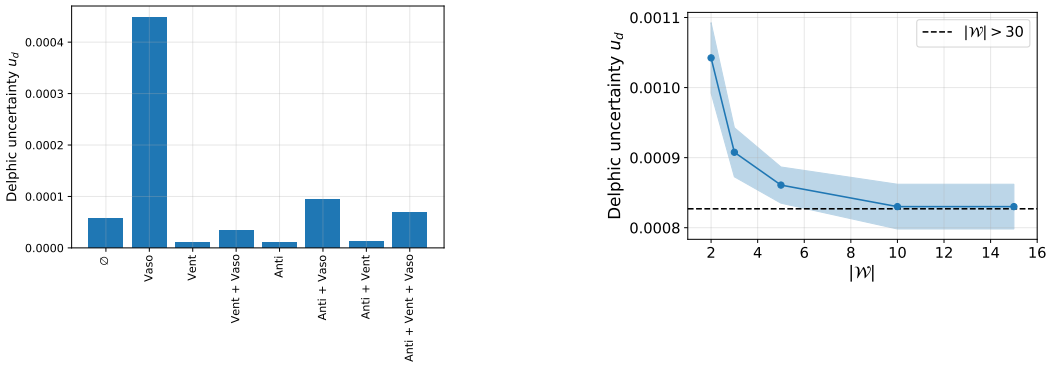
We contacted our local institution's ethics committee to enquire about the possible necessity of ethics approval for this experimental framework. We were informed that this was not considered necessary as the experts contribute to the validation of algorithms and are thus not themselves the subject of the research, and as the undertaking comes with minimal risks to those experts (anonymous data collection). Best practice was nonetheless observed, by providing participants with an information and consent letter to inform them of their rights and obligations, and of how their data is collected and used. Participants were asked to read and sign this letter before collecting their anonymous expert opinion.

Results in Figure 7 report the preference of clinicians for actions from either Delphic ORL or CQL or from behaviour cloning. We note their overall preference for the Delphic ORL policy in the confounded settings (high Γ). As more contextual information about the patient becomes available, however, and confounding is less marked (small Γ), physicians favour the behaviour cloning policy – closer to expected clinical practice.

E Ablations and Additional Results

E.1 Sepsis Environment

Ablation Study: Delphic Uncertainty. In Figure 9a, we find that delphic uncertainty is highest on the sepsis dataset when treatment involves vasopressors. By design of the simulation [48], this



(a) u_d as a function of a.

(b) u_d^π as a function of $|\mathcal{W}|$.

Figure 9: **Ablation Study: Delphic Uncertainty.** (a) Delphic uncertainty is highest under vasopressors in the sepsis environment, correctly identifying their confounded effect (Abbreviations: Vaso = Vasopressors, Anti = Antibiotics, Vent = Ventilation). (b) Empirically, only a small number of compatible worlds (for sepsis, $|\mathcal{W}| \approx 10$) is necessary to obtain an asymptotic estimate of u_d.

treatment is the only one for which patient evolution is confounded by the hidden diabetic status, which further supports the conclusion that delphic uncertainty captures model bias due to hidden confounding. In Figure 9b, we note that a only small number of world models (for sepsis, $|\mathcal{W}| \approx 10$) is necessary to obtain an estimate of delphic uncertainty consistent with a large number of world models. This motivates our practical choice to only consider a small set of world models to obtain a reasonable estimate of uncertainty for Delphic ORL, but warrants further theoretical work establishing guarantees and probability of correctness.

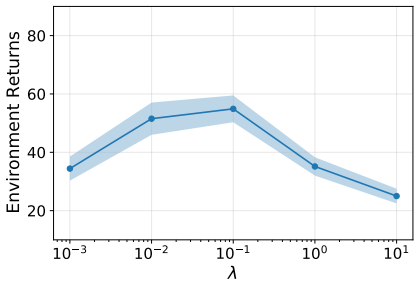


Figure 10: **Performance results as a function of hyperparameter** λ on the sepsis environment ($\Gamma = 46$).

Algorithm	Environment Returns
Online RL	67.8 ± 1.1
BC	38.5 ± 4.5
BCQ	18.5 ± 2.4
CQL	31.1 ± 3.5
Delphic ORL (u_d Threshold)	24.6 ± 3.4
Delphic BC (Weighting)	39.6 ± 4.1
Delphic ORL (Weighting)	44.7 ± 4.2
Delphic ORL (Algo. 2)	$\mathbf{54.9} \pm 4.6$

Table 4: **Performance of different pessimism methods** on the sepsis environment ($\Gamma = 46$).

Ablation Study: Delphic ORL. In Figure 10, we study the performance of Delphic ORL as a function of hyperparameter γ, interpolating between a naive implementation of Offline RL for very low values of γ (virtually no penalty) and an excessively pessimistic algorithm, where the confounding penalty overcomes any possible high-reward behaviour.

Next, Table 4 compares the performance of different approaches to implement pessimism with respect to delphic uncertainty. We find that our approach proposed in the main paper, based on penalising the target for the Bellman update, performs best in this experimental setting (sepsis dataset with $\Gamma = 46$). Weighting-based approaches also show promising performance (either matching or improving the performance of BC and CQL, respectively), which may be an avenue for further work and fine-tuning. Modifying the Bellman target to only include actions below a certain uncertainty threshold was however found to be excessively pessimistic, and degraded performance compared to the base CQL algorithm. Model-based Offline RL and Delphic ORL were not included as their performance was never found to improve over a random baseline policy. We hope this ablation study will motivate further work into the best possible approach to implement pessimism with respect to delphic uncertainty, to learn offline RL policies that are robust to hidden confounding bias.

E.2 Real-World Clinical Dataset

In this section, we provide additional evaluation metrics and investigations to understand the treatment strategies identified by the different algorithms considered, and in particular how Delphic ORL determines confounding-robust policies.

Table 5: **Difference in action choices from $\mathcal{D}_{test}$** across different algorithms (%). Our method learns a distinct policy from the doctors'. Mean and 95% CIs over 10 runs. Highest and overlapping values in bold.

Confounders $\mathcal{Z}$	BCQ	BC	CQL	**Delphic ORL**
All below	**32.2** $\pm$ 1.3	19.7 $\pm$ 1.1	**33.4** $\pm$ 0.9	**35.2** $\pm$ 1.5
{age}	**31.5** $\pm$ 1.3	12.8 $\pm$ 0.4	27.3 $\pm$ 0.3	**32.3** $\pm$ 0.5
{neuro. diag.}	31.1 $\pm$ 1.3	16.3 $\pm$ 1.0	**34.3** $\pm$ 1.3	30.6 $\pm$ 1.1
{gastro. diag.}	**27.1** $\pm$ 1.1	14.3 $\pm$ 0.9	**28.9** $\pm$ 1.1	**29.4** $\pm$ 1.3
{trauma}	**30.1** $\pm$ 1.5	12.8 $\pm$ 0.4	**24.2** $\pm$ 0.4	22.2 $\pm$ 0.7
{cardio. diag.}	28.7 $\pm$ 1.3	18.8 $\pm$ 1.2	**36.2** $\pm$ 1.3	29.6 $\pm$ 1.6
{endo. diag.}	**27.4** $\pm$ 1.3	13.5 $\pm$ 0.8	**27.3** $\pm$ 0.9	23.1 $\pm$ 0.9
{hemato. diag.}	**30.1** $\pm$ 1.5	12.4 $\pm$ 0.8	24.4 $\pm$ 1.1	23.6 $\pm$ 0.8
{weight}	**28.9** $\pm$ 1.3	13.2 $\pm$ 0.4	25.4 $\pm$ 0.6	23.6 $\pm$ 1.2
{sedation}	**30.5** $\pm$ 1.5	14.5 $\pm$ 0.7	25.1 $\pm$ 1.1	25.8 $\pm$ 1.0
{resp. diag.}	**27.7** $\pm$ 1.3	14.2 $\pm$ 0.6	28.5 $\pm$ 1.1	**25.2** $\pm$ 1.2
{intoxication}	**25.7** $\pm$ 1.1	12.6 $\pm$ 0.6	26.3 $\pm$ 0.6	23.1 $\pm$ 0.9
{surgical status}	**27.3** $\pm$ 1.3	14.3 $\pm$ 0.6	23.9 $\pm$ 0.8	22.1 $\pm$ 1.2
{ortho. diag.}	**25.6** $\pm$ 1.1	12.3 $\pm$ 0.6	24.1 $\pm$ 1.1	22.3 $\pm$ 1.0
{sepsis}	**26.1** $\pm$ 1.1	15.6 $\pm$ 0.8	23.5 $\pm$ 0.8	21.9 $\pm$ 1.2
$\emptyset$	**25.3** $\pm$ 0.9	12.2 $\pm$ 0.4	23.1 $\pm$ 0.8	21.7 $\pm$ 0.9

Table 5 provides a quantitative analysis of the disparities in action choices between different algorithms and the doctors' policy. As expected, behaviour cloning exhibits the closest resemblance to the doctors' treatment policy, which aligns with the characteristics of observational datasets. However, our proposed method outperforms behaviour cloning in terms of learning a distinct policy that deviates from the doctors' actions. These findings highlight the unique capabilities of our method in capturing important features and patterns beyond the direct imitation of doctors, enabling the model to make informed decisions that may differ from the observational data and potentially lead to improved treatment outcomes.

Following published recommendations on evaluating RL models in observational settings [17], we also analyse where policies differ most from the action choices in the observational dataset, and find that the policy learned by Delphic ORL diverges most at high disease severity (SOFA scores $\approx$ 15-20). In these cases, our policy appears to prescribe less fluids and vasopressors than in the data – which may be reasonable if unsure about possible adverse effects of an intervention. This relates to a comment received from one of the expert clinicians interviewed: "If I lack information about a patient [e.g. age, medical background and deliberately excluded variables], I would probably be more conservative with my treatment". Finally, we note a closer match to actions in the observational data at very high disease severity (SOFA score > 20), where negative rewards for *not* taking a therapeutic action outweighs potential confounding bias. Beyond this analysis, further insights could be gained by comparing interpretable representations of the trained policies [53], but we leave this as further work.