

---

# Interpretable Off-Policy Evaluation in Reinforcement Learning by Highlighting Influential Transitions

---

Omer Gottesman<sup>1</sup> Joseph Futoma<sup>1</sup> Yao Liu<sup>2</sup> Sonali Parbhoo<sup>1</sup> Leo Anthony-Celi<sup>3,4</sup> Emma Brunskill<sup>2</sup>  
Finale Doshi-Velez<sup>1</sup>

## Abstract

Off-policy evaluation in reinforcement learning offers the chance of using observational data to improve future outcomes in domains such as healthcare and education, but safe deployment in high stakes settings requires ways of assessing its validity. Traditional measures such as confidence intervals may be insufficient due to noise, limited data and confounding. In this paper we develop a method that could serve as a hybrid human-AI system, to enable human experts to analyze the validity of policy evaluation estimates. This is accomplished by highlighting observations in the data whose removal will have a large effect on the OPE estimate, and formulating a set of rules for choosing which ones to present to domain experts for validation. We develop methods to compute exactly the influence functions for fitted Q-evaluation with two different function classes: kernel-based and linear least squares. Experiments on medical simulations and real-world intensive care unit data demonstrate that our method can be used to identify limitations in the evaluation process and make evaluation more robust.

## 1. Introduction

Within reinforcement learning (RL), off-policy evaluation (OPE) is the task of estimating the value of a given evaluation policy, using data collected by interaction with the environment under a different behavior policy (Sutton & Barto, 2018; Precup, 2000). OPE is particularly valuable when interaction and experimentation with the environment is expensive, risky, or unethical—for example, in healthcare or with self-driving cars. However, despite recent interest and progress, state-of-the-art OPE methods still often fail to differentiate between obviously good and obviously bad

policies, e.g. in healthcare (Gottesman et al., 2018).

Most of the OPE literature focuses on sub-problems such as improving asymptotic sample efficiency or bounding the error on OPE estimators for the value of a policy. However, while these bounds are theoretically sound, they are often too conservative to be useful in practice (though see e.g. Thomas et al. (2019) for an exception). This is not surprising, as there is a theoretical limit to the statistical information contained in a given dataset, no matter which estimation technique is used. Furthermore, many of the common assumptions underlying these theoretical guarantees are usually not met in practice: observational healthcare data, for example, often contains many unobserved confounders (Gottesman et al., 2019a).

Given the limitations of OPE, we argue that in high stakes scenarios domain experts should be integrated into the evaluation process in order to provide useful actionable results. For example, senior clinicians may be able to provide insights that reduce our uncertainty of our value estimates. In this light, the explicit integration of expert knowledge into the OPE pipeline is a natural way for researchers to receive feedback and continually update their policies until one can make a responsible decision about whether to pursue gathering prospective data.

The question is then what information can humans provide that might help assess and potentially improve our confidence in an OPE estimate? In this work, we consider how human input could improve our confidence in the recently proposed OPE estimator, fitted Q-evaluation (FQE) (Le et al., 2019). We develop an efficient approach to identify the most influential transitions in the batch of observational data, that is, transitions whose removal would have large effects on the OPE estimate. By presenting these influential transitions to a domain expert and verifying that they are indeed representative of the data, we can increase our confidence that our estimated evaluation policy value is not dependent on outliers, confounded observations, or measurement errors. The main contributions of this work are:

- *Conceptual*: We develop a framework for using influence functions to interpret OPE, and discuss the types

---

<sup>1</sup>Harvard University <sup>2</sup>Stanford University <sup>3</sup>Massachusetts Institute of Technology <sup>4</sup>Beth Israel Deaconess Medical Center. Correspondence to: Omer Gottesman <gottesman@fas.harvard.edu>.

of questions which can be shared with domain experts to use their expertise in debugging OPE.

- *Technical*: We develop computationally efficient algorithms to compute the exact influence functions for two broad function classes in FQE: kernel-based functions and linear functions.
- *Empirical*: We demonstrate the potential benefits of influence analysis for interpreting OPE on a cancer simulator, and present results of analysis together with practicing clinicians of OPE for management of acute hypotension from a real intensive care unit (ICU) dataset.

## 2. Related work

The OPE problem in RL has been studied extensively. Works fall into three main categories: importance sampling (IS) (e.g. Precup (2000); Jiang & Li (2015)), model-based (often referred to as the direct method) (e.g. Hanna et al. (2017); Gottesman et al. (2019b)), and value-based (e.g. Le et al. (2019)). Some of these works provide bounds on the estimation errors (e.g. Thomas et al. (2015); Dann et al. (2018)). We emphasize, however, that for most real-world applications these bounds are either too conservative to be useful or rely on assumptions which are usually violated.

While there has been considerable recent progress in interpretable machine learning and machine learning with humans in the loop (e.g. Tamuz et al. (2011); Lage et al. (2018)), to our knowledge, there has been little work that considers human interaction in the context of OPE. Oberst & Sontag (2019) proposed framing the OPE problem as a structural causal model, which enabled them to identify trajectories where the predicted counterfactual trajectories under an evaluation policy differs substantially from the observed data collected under the behavior policy. However, that work does not give guidance on what part of the trajectory might require closer scrutiny, nor can it use human input for additional refinement.

Finally, the notion of influence that we use throughout this work has a long history in statistics as a technique for evaluating the robustness of estimators (Cook & Weisberg, 1980). Recently, an approximate version of influence for complex black-box models was presented in Koh & Liang (2017), and they demonstrated how influence functions can make machine learning methods more interpretable. In the context of optimal control and RL, influence functions were first introduced by Munos & Moore (2002) to aid in online optimization of policies. However, their definition of influence as a change in the value function caused by perturbations of the reward at a specific state is quite different from ours.

## 3. Background

### 3.1. Notation

A Markov Decision Process (MDP) is a tuple  $\langle \mathcal{X}, \mathcal{A}, P_T, P_R, P_0, \gamma \rangle$ , where  $\mathcal{X}$ ,  $\mathcal{A}$  and  $\gamma$  are the state space, action space, and the discount factor, respectively. The next state transition and reward distributions are given by  $P_T(\cdot|x, a)$  and  $P_R(\cdot|x, a)$  respectively, and  $P_0(x)$  is the initial state distribution. The state and action spaces could be either discrete or continuous, and the transition and reward functions may be either stochastic or deterministic.

A dataset is composed of a set of  $N$  observed transitions  $\mathcal{D} = \{(x^{(n)}, a^{(n)}, r^{(n)}, x'^{(n)})\}_{n=1}^N$ , and we use  $\tau^{(n)}$  to denote a single transition. The subset  $\mathcal{D}_0 \subseteq \mathcal{D}$  denotes initial transitions from which  $P_0$  can be estimated. Note that although we treat all data points as observed transitions, in most practical applications data is collected in the form of trajectories rather than individual transitions.

A policy is a function  $\pi : (\mathcal{X}, \mathcal{A}) \rightarrow [0, 1]$  that gives the probability of taking each action at a given state ( $\sum_{a \in \mathcal{A}} \pi(a|x) = 1$ ). The value of a policy is the expected return collected by following the policy,  $v^\pi := \mathbb{E}[g_T | a_t \sim \pi]$ , where actions are chosen according to  $\pi$ , expectations are taken with respect to the MDP, and  $g_T := \sum_{t=0}^T \gamma^t r_t$  denotes the total return, or sum of discounted rewards. The state-action value function  $q^\pi(x, a)$  is the expected return for taking action  $a$  at state  $x$ , and afterwards following  $\pi$  in selecting future actions. The goal of off-policy evaluation is to estimate the value of an *evaluation* policy,  $\pi_e$ , using data collected using a different *behavior* policy  $\pi_b$ . In this work, we are only interested in estimating  $v^{\pi_e}$  and  $q^{\pi_e}$ , and will therefore drop the superscript for brevity. We will also limit ourselves to deterministic evaluation policies.

For the purpose of kernel-based value function approximation, we define a distance metric,  $d((x^{(i)}, a^{(i)}), (x^{(j)}, a^{(j)}))$  over  $\mathcal{X} \times \mathcal{A}$ . In this work, for discrete action spaces, we will assume  $d((x^{(i)}, a^{(i)}), (x^{(j)}, a^{(j)})) = \infty$  when  $a^{(i)} \neq a^{(j)}$ , but this is not required for any of the derivations.

### 3.2. Fitted Q-Evaluation

Fitted Q-Evaluation (Le et al., 2019) can be thought of as dynamic programming on an observational dataset to compute the value of a given evaluation policy. It is similar to the more well-known fitted Q-iteration method (FQI) (Ernst et al., 2005), except it is performed offline on observational data, and the target is used for evaluation of a given policy rather than for optimization. FQE performs a sequence of supervised learning steps where the inputs are state-action pairs, and the targets at each iteration are given by  $y_i(x, a) = r + \gamma \hat{q}_{i-1}(x', \pi_e(x'))$ , where  $\hat{q}_{i-1}(x, a)$  is the estimator (from a function class  $\mathcal{F}$ ) that best estimates

$y_{i-1}(x, a)$ . For more information, see Le et al. (2019).

## 4. OPE diagnostics using influence functions

### 4.1. Definition of the influence

We aim to make OPE interpretable and easy to debug by identifying transitions in the data which are highly influential on the estimated policy value. We define the *total influence* of transition  $\tau^{(j)}$  as the change in the value estimate if  $\tau^{(j)}$  was removed:

$$I_j \equiv \hat{v}_{-j} - \hat{v}, \quad (1)$$

where  $\hat{v}_{-j}$  is the value estimate using the same dataset after removal of  $\tau^{(j)}$ . In general, for any function of the data  $f(\mathcal{D})$  we will use  $f(\mathcal{D}_{-j}) \equiv f_{-j}$  to denote the value of  $f$  computed for the dataset after removal of  $\tau^{(j)}$ .

Another quantity of interest is the change in the estimated value of  $q(x^{(i)}, a^{(i)})$  as a result of removing  $\tau^{(j)}$ , which we call the *individual influence*:

$$I_{i,j} \equiv \hat{q}_{-j}(x^{(i)}, a^{(i)}) - \hat{q}(x^{(i)}, a^{(i)}). \quad (2)$$

The total influence of  $\tau^{(j)}$  can be computed by averaging its individual influences over the set  $\mathcal{D}_0^*$  of all initial state-action transitions in which  $a = \pi_e(x)$ :

$$I_j = \frac{1}{|\mathcal{D}_0^*|} \sum_{i \in \mathcal{D}_0^*} I_{i,j}. \quad (3)$$

As we are interested in the robustness of our evaluation, we can normalize the absolute value of the influence of  $\tau^{(j)}$  by the estimated value of the policy to provide a more intuitive notion of overall importance:

$$\tilde{I}_j \equiv \frac{|I_j|}{|\hat{v}|}. \quad (4)$$

### 4.2. Diagnosing OPE estimation

With the above definitions of influence functions, we now formulate and discuss guidelines for diagnosing the OPE process for potential problems.

**No influential transitions: OPE appears reliable.** As a first diagnostic, we check that none of the transitions influence the OPE estimate by more than a specified influence threshold  $\tilde{I}_C$ , i.e. for all  $j$  we have  $\tilde{I}_j \leq \tilde{I}_C$ . In such a case we would output that, to the extent that low influences suggests the OPE is stable, the evaluation appears reliable. That said, we emphasize that our proposed method for evaluating OPE methods is not exhaustive, and there could be many other ways in which OPE could fail.

**Influential transitions: a human can help.** When there are several influential transitions in the data (defined as transitions whose influence is larger than  $\tilde{I}_C$ ), we present them to domain experts to determine whether they are representative, that is, taking action  $a$  in state  $x$  is likely to result in transition to  $x'$ . If the domain experts can validate all influential transitions, we can still have some confidence in the validity of the OPE. If any influential transitions are flagged as unrepresentative or artefacts, we have several options: (1) Declare the OPE as unreliable; (2) Remove the suspect influential transitions from the data and recompute the OPE; (3) Caveat the OPE results as valid only for a subset of initial states that do not rely on that problematic transition.

In situations where there is a large number of influential transitions, manual review by experts may be infeasible. As such, it is necessary to present as few transitions as possible while still presenting enough to ensure that any potential artefacts in the data and/or the OPE process are accounted for. In practice, we find it is common to observe a sequence of influential transitions where removing any single transition has the same effect as removing the entire sequence. An example of this is shown schematically in Figure 1. An entire sequence marked in blue and red leads to a region of high reward, and so all transitions in that sequence will have high influence. The whole influential sequence appears very different from the rest of the data, and a domain expert might flag it as an outlier to be removed. However, we can present the expert with only the red transition and capture the influence of the blue transitions as well, reducing the number of suspect examples to be manually reviewed.

**Influential transitions: policy is unevaluable.** When an influential transition,  $\tau^{(j)}$ , has no nearest neighbors to  $(x'^{(j)}, \pi_e(x'^{(j)}))$ , we can determine that the evaluation policy cannot be evaluated, even without review by a domain expert. This claim is a result of the fact that such a situation represents reliance of the OPE on transitions for which there is no overlap between the actions observed in the data and the evaluation policy. However, while the evaluation policy is not evaluable, the influential “dead-end” transitions may still inform experts of what data is required for evaluation to be feasible.

It should be noted that the applicability of the diagnostics methods discussed above may change depending on whether the FQE function class is parametric or nonparametric. All function classes lend themselves to highlighting of highly influential transitions. However, the notion of stringing together sequences of neighbors, or looking for red flags in the form of influential transitions with no neighbors to their  $(x', \pi_e(x'))$  state action pairs only makes sense for nonparametric models. In the case of parametric models, the notion of neighbors is less important as the influence of removing

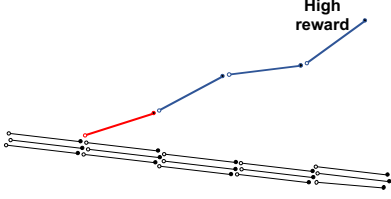


Figure 1. **Schematic of an influential sequence.** All transitions in the sequence leading to a high reward have high influence, but flagging just the red transition for inspection will capture the influence of the blue ones as well.

a transition manifests as a change to the learned parameters which affects the value estimates for the entire domain simultaneously. In contrast, for nonparametric methods, removing a transition locally changes the value of neighboring transitions and propagates through the entire domain through the sequential nature of the environment. While we derive efficient ways to compute the influence for both parametric and nonparametric function classes, in the empirical section of this paper we present results for nonparametric kernel-based estimators to demonstrate all diagnostics.

## 5. Efficient computation of influence functions

A key technical challenge in performing the proposed influence analysis in OPE is computing the influences efficiently. The brute-force approach of removing a transition and re-computing the OPE estimate is clearly infeasible for all but tiny problems, as it requires refitting  $N$  models. The computation of influences in RL is also significantly more challenging than in static one-step prediction tasks, as a change in the value of one state has a ripple effect on all other states that are possible to reach from it. We describe computationally efficient methods to compute the influence functions in two classes of FQE: kernel-based, and linear least squares. Unlike previous works (e.g. (Koh & Liang, 2017)) that approximate the influence function for a broad class of black-box functions, we provide closed-form, analytic solutions for the exact influence function in two widely used white-box function classes.

### 5.1. Kernel-Based FQE

In kernel based FQE, the function class we choose for estimating the value function of  $\pi_e$  at a point in state-action space is based on similar observations within that space. For simplicity, in the main body of this work we estimate the value function as an average of all its neighbors within a ball of radius  $R$ , i.e.

$$\hat{q}(x, a) = \frac{1}{N_{(x,a)}} \sum_i \hat{q}(x^{(i)}, a^{(i)}) \quad (5)$$

where the summation is performed over all  $(x^{(i)}, a^{(i)})$  such that  $d((x^{(i)}, a^{(i)}), (x, a)) < R$  and  $N_{(x,a)}$  is the number of such points. Extension to general kernel functions is straightforward. We introduce a matrix formulation for performing FQE which allows for efficient computation of the influence functions.

### Matrix formulation of nearest-neighbors based FQE.

We define  $\Delta_{ij}$  as the event that the starting state-action of  $\tau^{(j)}$  is a neighbor of the starting state-action of  $\tau^{(i)}$ , i.e.  $d((x^{(i)}, a^{(i)}), (x^{(j)}, a^{(j)})) < R$ . Similarly, we define  $\Delta'_{ij}$  as the event that the starting state-action of  $\tau^{(j)}$  is a neighbor of the next-state and corresponding  $\pi_e$  action of  $\tau^{(i)}$ , i.e.  $d((x'^{(i)}, \pi(x'^{(i)})), (x^{(j)}, a^{(j)})) < R$ . We also define the counts for numbers of neighbors of trajectories as  $N_i = \sum_{j=1}^N \mathbb{I}(\Delta_{ij})$  and  $N_{i'} = \sum_{j=1}^N \mathbb{I}(\Delta'_{ij})$ , where  $\mathbb{I}(e)$  is the indicator function.

To perform nearest-neighbors FQE using matrix multiplications, we first construct two nearest-neighbors matrices: one for the neighbors of all state-action pairs, and one for the neighbors of all state-action pairs with pairs of next-states and subsequent actions under  $\pi_e$ . Formally:

$$\mathbf{M}_{ij} = \frac{\mathbb{I}(\Delta_{ij})}{N_i}; \quad \mathbf{M}'_{ij} = \frac{\mathbb{I}(\Delta'_{ij})}{N_{i'}}. \quad (6)$$

The  $N \times N$  matrices  $\mathbf{M}$  and  $\mathbf{M}'$  can be easily computed from the data, and are used to compute the value function for all state-action pairs using the following proposition, the proof of which is given in Appendix A.1.

**Proposition 1.** *For all transitions in the dataset, the values for corresponding state-action pairs are given by*

$$\hat{q}'_t = \left( \sum_{t'=1}^t \gamma^{t'-1} \mathbf{M}'^{t'} \right) \mathbf{r} \equiv \Phi'_t \mathbf{r} \quad (7)$$

$$\hat{q}_t = \mathbf{M} \left( \sum_{t'=1}^t (\gamma \mathbf{M}')^{t'-1} \right) \mathbf{r} \equiv \Phi_t \mathbf{r}. \quad (8)$$

where  $\hat{q}'_{t,i}$  and  $\hat{q}_{t,i}$  are the estimated policy values at  $(x'^{(i)}, \pi_e(x'^{(i)}))$  and  $(x^{(i)}, a^{(i)})$ , respectively, for  $\tau^{(i)}$ .

In future derivations, we will drop the time dependence of  $\Phi$  and  $\hat{q}$  on  $t$ . This is justified when there are well defined ends of trajectories with no nearest neighbors (or equivalently, trajectories end in an absorbing state), and the number of iterations in the FQE is larger than the longest trajectory.

**Influence function computation.** Removal of a transition  $\tau^{(j)}$  from the dataset can affect  $\hat{q}_i$  in two ways. First,  $\hat{q}_i$  is a mean over all of its neighbors, indexed by  $k$ , of  $r^{(k)} + \gamma \hat{q}'_k$ . Thus if  $(x^{(j)}, a^{(j)})$  is one of the  $M_{ij}^{-1}$  neighbors of  $(x^{(i)}, a^{(i)})$ , removing it from the dataset will change the value of  $\hat{q}_i$  by  $\frac{\hat{q}_i - (r^{(j)} + \gamma \hat{q}'_j)}{M_{ij}^{-1} - 1}$ . The special case of  $M_{ij}^{-1} = 1$

does not pose a problem in the denominator, as given that  $i \neq j$  and every transition is a neighbor of itself, if  $(x^{(j)}, a^{(j)})$  is a neighbor of  $(x^{(i)}, a^{(i)})$ , then  $M_{ij}^{-1} \geq 2$ .

The second way in which removing  $\tau^{(j)}$  influences  $\hat{q}_i$  is through its effect on intermediary transitions. Removal of  $\tau^{(j)}$  changes the estimated value of  $\hat{q}'_k$ , of all  $(x^{(k)}, \pi_e(x^{(k)}))$  that  $(x^{(j)}, a^{(j)})$  is a neighbor of by  $\frac{\hat{q}'_k - (r^{(j)} + \gamma \hat{q}'_j)}{M'_{kj} - 1}$ . Multiplying this difference by  $\gamma$  yields the difference in  $\hat{q}_k$  due to removal of  $\tau^{(j)}$ . A change in the value of  $\hat{q}_k$  is identical in its effect on the value estimation to changing  $r^{(k)}$ , a change which is mediated to  $\hat{q}_i$  through  $\Phi_{ik}$ . In the special case that  $(x^{(j)}, a^{(j)})$  is the only neighbor of  $(x^{(k)}, \pi_e(x^{(k)}))$ , the value estimate  $\hat{q}'_k$  changes from  $\hat{q}_j$  to zero.

Combining the two ways in which removal of  $\tau^{(j)}$  changes the estimated value  $\hat{q}_i$  yields the individual influence:

$$I_{i,j} = \mathbb{I}(\Delta_{ij}) \frac{\hat{q}_i - (r^{(j)} + \gamma \hat{q}'_j)}{M_{ij}^{-1} - 1} + \sum_{k: \Delta_{kj}} I_{i,j}^{(k)}, \quad (9)$$

where we define

$$I_{i,j}^{(k)} = \begin{cases} \gamma \Phi_{ik} \frac{\hat{q}'_k - (r^{(j)} + \gamma \hat{q}'_j)}{M'_{kj} - 1} & M'_{kj} > 1 \\ \gamma \Phi_{ik} \hat{q}_j & M'_{kj} = 1. \end{cases} \quad (10)$$

**Computational complexity.** The matrix formulation of kernel based FQE allows us to compute an individual influence in constant time, making influence analysis of the entire dataset possible in  $\mathcal{O}(N|D_0^*|)$  time. Furthermore, the sparsity of  $\mathbf{M}$  and  $\mathbf{M}'$  allows the FQE itself to be done in  $\mathcal{O}(N^2T)$ . See Appendix A.2 for a full discussion.

## 5.2. Linear Least Squares FQE

In linear least squares FQE, the policy value function  $\hat{q}(x, a)$  is approximated by a linear function  $\hat{q}(x, a) = \boldsymbol{\psi}(x, a)^\top \mathbf{w}$  where  $\boldsymbol{\psi}(x, a)$  is a  $D$ -dimensional feature vector for a state-action pair. Let  $\boldsymbol{\Psi} \in \mathbb{R}^{N \times D}$  be the sample matrix of  $\boldsymbol{\psi}(x, a)$ . Define vector  $\boldsymbol{\psi}_\pi(x) = \gamma \boldsymbol{\psi}(x, \pi_e(x))$  and let  $\boldsymbol{\Psi}_p \in \mathbb{R}^{N \times D}$  be the sample matrix of  $\boldsymbol{\psi}_\pi(x')$ . The least-squares solution of  $\mathbf{w}$  is  $(\boldsymbol{\Psi}^\top \boldsymbol{\Psi} - \gamma \boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p)^{-1} \boldsymbol{\Psi}^\top \mathbf{r}$  (See Appendix 2 for full derivation).

Let  $\mathbf{w}_{-j}$  be the solution of linear least squares FQE after removing  $\tau^{(j)}$ , and  $\boldsymbol{\Psi}_{-j}$ ,  $\mathbf{r}_{-j}$ , and  $\boldsymbol{\Psi}_{p,-j}$  be the corresponding matrices and vectors without the  $\tau^{(j)}$ . Then,  $\mathbf{w}_{-j} = (\boldsymbol{\Psi}_{-j}^\top \boldsymbol{\Psi}_{-j} - \gamma \boldsymbol{\Psi}_{p,-j}^\top \boldsymbol{\Psi}_{p,-j})^{-1} \boldsymbol{\Psi}_{-j}^\top \mathbf{r}_{-j}$ . The key challenge of computing the influence function is computing  $\mathbf{w}_{-j}$  in an efficient manner that avoids recomputing a costly matrix inverse for each  $j$ . Let  $\mathbf{C}_{-j} = (\boldsymbol{\Psi}_{-j}^\top \boldsymbol{\Psi}_{-j} - \gamma \boldsymbol{\Psi}_{p,-j}^\top \boldsymbol{\Psi}_{p,-j})$ .

and  $\mathbf{C} = (\boldsymbol{\Psi}^\top \boldsymbol{\Psi} - \gamma \boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p)$ . We compute  $\mathbf{w}_{-j}$  as follows:

$$\mathbf{B}_j \leftarrow \mathbf{C}^{-1} + \frac{\mathbf{C}^{-1} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top \mathbf{C}^{-1}}{1 - \boldsymbol{\psi}_j^\top \mathbf{C}^{-1} \boldsymbol{\psi}_j} \quad (11)$$

$$(\mathbf{C}_{-j})^{-1} \leftarrow \mathbf{B}_j - \frac{\gamma \mathbf{B}_j \boldsymbol{\psi}_j \boldsymbol{\psi}_{\pi,j}^\top \mathbf{B}_j}{1 + \gamma \boldsymbol{\psi}_{p,j}^\top \mathbf{B}_j \boldsymbol{\psi}_j} \quad (12)$$

$$\mathbf{w}_{-j} \leftarrow (\mathbf{C}_{-j})^{-1} (\boldsymbol{\Psi}^\top \mathbf{r} - r^{(j)} \boldsymbol{\psi}_j) \quad (13)$$

The proof of correctness is in Proposition 3 in Appendix B. The individual influence function is then simply:

$$I_{i,j} = \boldsymbol{\psi}(s^{(i)}, a^{(i)})^\top (\mathbf{w}_{-j} - \mathbf{w}). \quad (14)$$

**Computational complexity.** The bottleneck of computing  $\mathbf{w}_{-j}$  is the matrix multiplication of  $D \times D$  matrices which takes at most  $\mathcal{O}(D^3)$ . All the other matrix multiplications involving size  $N$ , e.g.  $\boldsymbol{\Psi}^\top \mathbf{r}$ , do not depend on  $j$  and could be cached from the original OPE. Thus, the overall complexity for computing  $I_{i,j}$  for all  $i$  and  $j$  is  $\mathcal{O}(ND^3)$ . Assuming  $N > D$ , the complexity of the original OPE algorithm is  $\mathcal{O}(ND^2)$ , where the bottleneck is computing  $\boldsymbol{\Psi}^\top \boldsymbol{\Psi}$ .

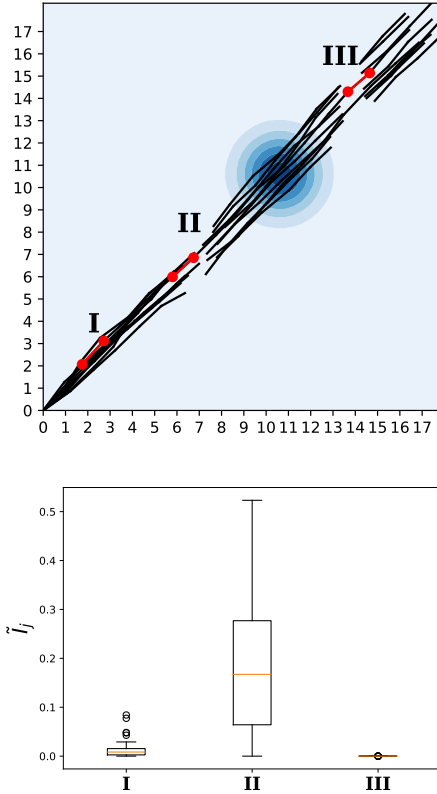
## 6. Illustration of influence functions in a sequential setting

We now demonstrate and give intuition for how the influence behaves in an RL setting. For the demonstrations and experiments presented throughout the rest of the paper we use the kernel-based FQE method.

Several factors determine the influence of a transition. For transitions to be influential they must have actions which are possible under the evaluation policy and form links in sequences which result in returns different than the expected value. Furthermore, transitions will be more influential the less neighbors they have.

To demonstrate this intuition we present in Figure 2 trajectories from a 2D continuous navigational domain. The agent starts at the origin and takes noisy steps of length 1 at  $45^\circ$  to the axes. The reward for a given transition is a function of the state and has the shape of a Gaussian centered along the approximate path of the agent, represented as the background heat map in Figure 2 (top), where observed transitions are drawn as black line segments. Because distances for the FQE are computed in the state-action space, in this example all actions in the data are the same to allow for distances to be visualized in 2D.

To illustrate how influence is larger for transitions with few neighbors, we removed most of the transitions in two regions (denoted II and III), and compared the distribution of influences in these regions with influences in a data dense



**Figure 2. Conceptual demonstration on a 2D domain.** For transitions in the data to have high influence, they must agree with the evaluation policy and lead to rewarding regions in the state-action space. Additionally, the influence of transitions decreases with the number of their close neighbors.

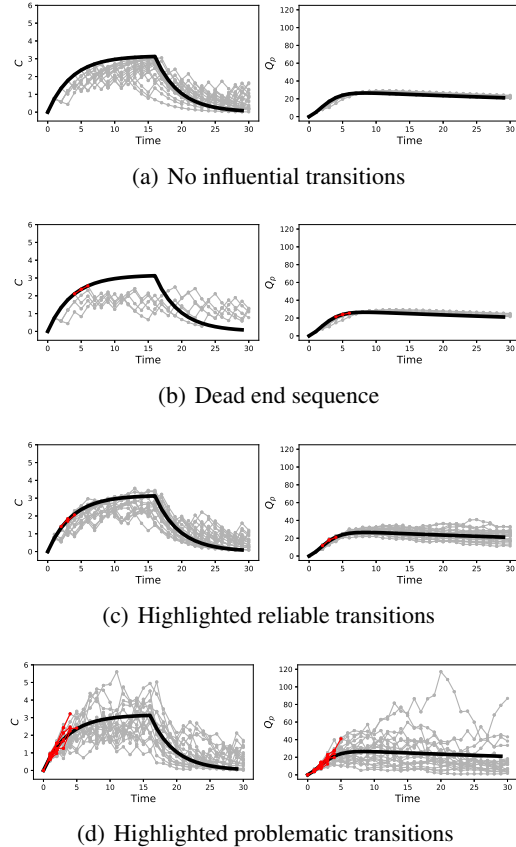
region (denoted I). Figure 2 (bottom) shows the distribution over 200 experiments (in each experiment, new data is generated) of the influences of transitions in the different regions. The influence is much higher for transitions in sparse regions with few neighbors, as can be seen by comparing the distributions in regions I and II. This is a desired property, as in analysis of the OPE process, we’d like to be able to present domain experts with transitions that have few neighbors where the sampling variance of a particular transition could have large effect on evaluation.

In region III, despite the fact that the observations examined also have very few neighbors, their influence is extremely low, as they don’t lead to any regions where rewards are gained by the agent.

## 7. Experiments

### 7.1. Medical simulators

To demonstrate the different ways in which influence analysis can allow domain experts to either increase our con-



**Figure 3. Influence analysis for simulated cancer data.** Analysis of synthetic cancer simulations demonstrates how influence analysis can differentiate between different diagnostics of the OPE process.

fidence in the validity of OPE or identify instances where they are invalid, we first present results on a simulator of cancer dynamics. The 4 dimensional states of the simulator approximate the dynamics of tumor growth, with actions consisting administration of chemotherapy at each timestep representing one month. For more details see Ribba et al. (2012).

In Figure 3 we present four cases in which we attempt to evaluate the policy of treating a patient for 15 months and then discontinuing chemotherapy until the end of treatment at 30 months. Each subplot in Figure 3 shows two of the four state variables as a function of time, under different conditions which might make evaluation more difficult, such as difference in behavior policy or stochasticity in the environment. The heavy black line represents the expectation of each state dimension at each time-step under the evaluation policy, while the grey lines represent observed transitions under the behavior policy which is  $\epsilon$ -greedy with respect to the evaluation policy. In all figures, we highlight in red all influential transitions our method would have highlighted

for review by domain experts ( $\tilde{\epsilon}_c = 0.05$ ).

**Case 1: OPE seems reliable.** Figure 3(a) represents a typical example where the OPE can easily be trusted. Despite the large difference between the evaluation and behavior policy ( $\epsilon = 0.3$ ), enough trajectories have been observed in the data to allow for proper evaluation, and no transition is flagged as being too influential. The value estimation error in this example is less than 1% and our method correctly labels this dataset as reliable.

**Case 2: Unevaluatable.** Figure 3(b) is similar in experimental conditions to (a) ( $\epsilon = 0.3$  and deterministic transitions), but with less collected data, so that the observations needed to properly estimate the dynamics are not in the data. This can be seen by the lack of overlap between the observed transitions and the expected trajectory, and results in a 38% value estimation error. In real life we will not know what the expected trajectory under the evaluation policy looks like, and therefore will not be able to make the comparison and detect the lack of overlap between transitions under the evaluation and behavior policies. However, our method highlights a very influential sequence which terminates at a dead-end, and thus will correctly flag this dataset as not sufficient for evaluation. Our method in this case is confident enough to dismiss the results of evaluation without need for domain experts, but can still inform experts on what type of data is lacking in order for evaluation to be feasible.

**Case 3: Humans might help.** In Figures 3(c-d),  $\epsilon = 0.3$ , but the dynamics have different levels of stochasticity. The less stochastic dynamics in 3(c) allow for relatively accurate evaluation (8% error) but our method identifies several influential transitions which must be presented to a domain expert. These transitions lie on the expected trajectory, and thus a clinician would verify that they represent a typical response of a patient to treatment. This is an example in which our method would allow a domain expert to verify the validity of the evaluation by examining the flagged influential transitions.

Conversely, in 3(d) some extreme outliers lead to a large estimation error (23% error). The influential transitions identified by our method are exactly those which start close to the expected trajectory but deviate significantly from the expected dynamics. A domain expert presented with these transitions would easily be able to note that the OPE heavily relies on atypical patients and rightly dismiss the validity of evaluation.

To summarize this section, we demonstrated that analysis of influences can both validate or invalidate the evaluation without need for domain experts, and in intermediate cases present domain experts with the correct queries required to gain confidence in the evaluation results or dismiss them.

## 7.2. Analysis of real ICU data - MIMIC III

To show how influence analysis can help debug OPE for a challenging healthcare task, we consider the management of acutely hypotensive patients in the ICU. Hypotension is associated with high morbidity and mortality (Jones et al., 2006), but management of these patients is not standardized as ICU patients are heterogeneous. Within critical care, there is scant high-quality evidence from randomized controlled trials to inform treatment guidelines (de Grooth et al., 2018; Girbes & de Grooth, 2019), which provides an opportunity for RL to help learn better treatment strategies. In collaboration with an intensivist, we use influence analysis to identify potential artefacts when performing OPE on a clinical dataset of acutely hypotensive patients.

**Data and evaluation policy.** Our data source is a subset of the publicly available MIMIC-III dataset (Johnson et al., 2016). See Appendix C for full details of the data preprocessing. Our final dataset consists of 346 patient trajectories (6777 transitions) for learning a policy and another 346 trajectories (6863 transitions) for evaluation of the policy via OPE and influence analysis.

Our state space consists of 29 relevant clinical variables, summarizing current physiological condition and past actions. The two main treatments for hypotension are administration of an intravenous (IV) fluid bolus or initiation of vasopressors. We bin doses of each treatment into 4 categories for "none", "low", "medium" and "high", so that the full action space consists of 16 discrete actions. Each reward is a function of the next blood pressure (MAP) and takes values in  $[-1, 0]$ . As an evaluation policy, we use the most common action of a state's 50 nearest neighbors. This is setup is equivalent to constructing a decision assistance tool for clinicians by recommending the common practice action for patients, and using OPE combined with influence analysis to estimate the efficacy of such a tool. See Appendix C for more details on how we setup the RL problem formulation, and for the kernel function used to compute nearest-neighbors.

**Presenting queries to a practicing intensivist.** Running influence analysis flags 6 influential transitions that have high influence, which we define as a change of 5% or more on the final value estimate. We show 2 of these transitions in Figure 4 and the rest in Appendix D. While this analysis highlights individual transitions, our results figures display additional context before and after the suspect transition to help the clinician understand what might be going on.

In Figure 4, each column shows a transition flagged by influence analysis. The top two rows show actions taken (actual treatments in the top row and binned actions in the second row). The remaining three rows show the most important

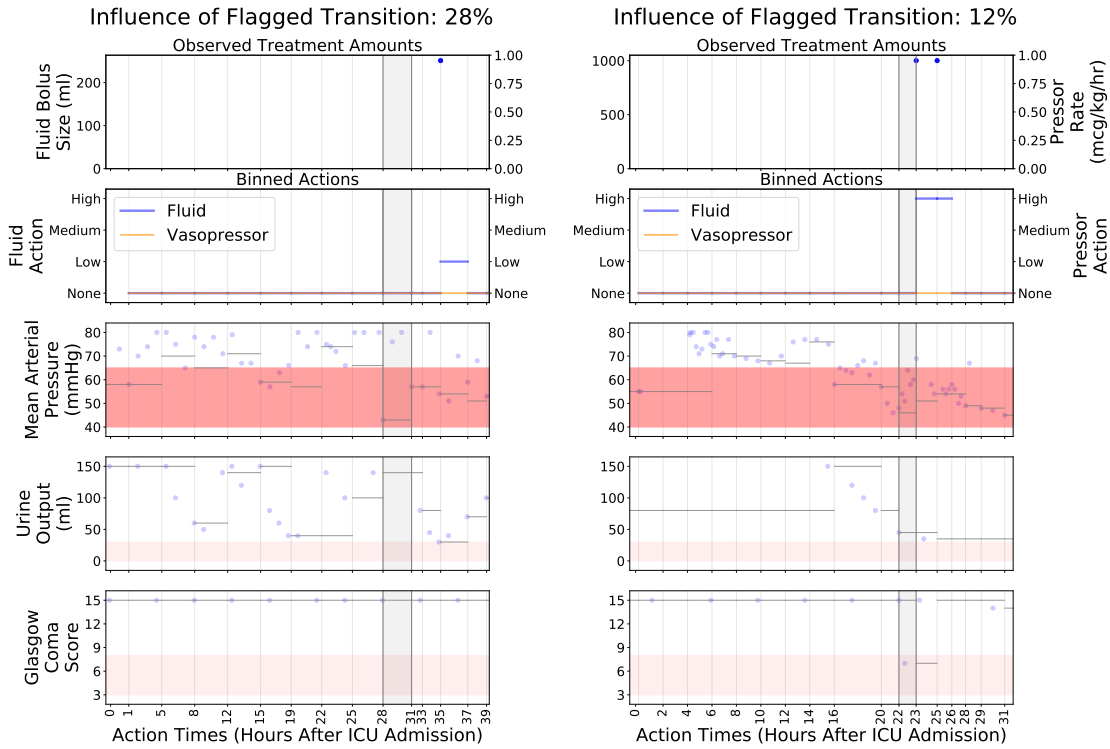


Figure 4. Influence analysis on our real-world dataset discovered six transitions in the evaluation dataset that were especially influential on our OPE. We display two of them in this figure, see Appendix D for the remaining four.

state variables that inform the clinicians’ decisions: blood pressure (MAP), urine output, and level of consciousness (GCS). For these three variables, the abnormal range is shaded in red, where the blood pressure shading is darker highlighting its direct relationship with the reward. Vertical grey lines represent timesteps, and the highlighted influential transition is shaded in grey.

**Outcome: Identifying and removing an influential, buggy measurement.** The two transitions in Figure 4 highlight potential problems in the dataset that have a large effect on our final OPE estimate. In the first transition (left), a large drop in blood pressure is observed at the starting time of this transition, potentially leaving the patient in a dangerous hypotensive state. Surprisingly, the patient received no treatment, and this unusual transition has a 29% influence on the OPE estimate. Given additional context just before and after the transition, though, it was clear to the clinician that this was due to a single low measurement in a sequence that was previously stable. Coupled with a stable GCS (patient was conscious and alert) and a normal urine output, the intensivist easily determined the single low MAP was likely either a measurement error or a clinically insignificant transient episode of hypotension. After correcting the outlier MAP measurement to its most recent normal value (80mmHg) and then rerunning FQE and the influence

analysis, the transition no longer has high influence and was not flagged.

**Outcome: Identifying and correcting a temporal misalignment.** The second highlighted transition (right) features a sudden drop in GCS and worsening MAP values, indicating a sudden deterioration of the patient’s state, but treatment is not administered until the next timestep. The intensivist attributed this finding to a time stamp recording error. Again, influence analysis identified an inconsistency in the original data which had undue impact on evaluation. After correcting the inconsistency by shifting the two fluid treatments back by one timestep each, we found that the transition no longer had high influence and was not flagged.

## 8. Discussion

A key aim of this paper is to formulate a framework for using domain expertise to help in evaluating the trustworthiness of OPE methods for noisy and confounded observational data. The motivation for this research direction is the intersection of two realities: for messy real-world applications, the data itself might never be enough; and domain experts will always need to be involved in the integration of decision support tools in the wild, so we should incorporate their expertise into the evaluation process. We showcased influence



analysis as one way of performing this task for value-based OPE, but emphasize that such measures can and should be incorporated into other OPE methods as well. For example, importance sampling weights offer a straightforward way of highlighting important entire trajectories for IS based techniques, and the dynamics learned by models in model-based OPE can be tested for their agreement with expert intuition.

We stress that research to integrate human input into OPE methods to increase their reliability complements, and does not replace, the approaches for estimating error bounds and uncertainties over the errors of OPE estimates. The fact that traditional theoretical error bounds rely so heavily on assumptions which are generally impossible to verify from the data alone highlights the need for other techniques for gauging to what extent these assumptions hold.

## References

- Agniel, D., Kohane, I. S., and Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361:k1479, 2018.
- Asfar, P., Meziani, F., Hamel, J.-F., Grelon, F., Megarbane, B., Anguel, N., Mira, J.-P., Dequin, P.-F., Gergaud, S., Weiss, N., et al. High versus low blood-pressure target in patients with septic shock. *N Engl J Med*, 370:1583–1593, 2014.
- Cook, R. D. and Weisberg, S. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.
- Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*, 2018.
- de Grooth, H.-J., Postema, J., Loer, S. A., Parienti, J.-J., Oudemans-van Straaten, H. M., and Girbes, A. R. Unexplained mortality differences between septic shock trials: a systematic analysis of population characteristics and control-group mortality rates. *Intensive care medicine*, 44(3):311–322, 2018.
- Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.
- Girbes, A. R. J. and de Grooth, H.-J. Time to stop randomized and large pragmatic trials for intensive care medicine syndromes: the case of sepsis and acute respiratory distress syndrome. *Journal of Thoracic Disease*, 12(S1), 2019. ISSN 2077-6624. URL <http://jtd.amegroups.com/article/view/33636>.
- Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.
- Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nat Med*, 25(1):16–18, 2019a.
- Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., and Doshi-Velez, F. Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pp. 2366–2375, 2019b.
- Hanna, J. P., Stone, P., and Niekum, S. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.
- Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- Jones, A. E., Yiannibas, V., Johnson, C., and Kline, J. A. Emergency department hypotension predicts sudden unexpected in-hospital mortality: a prospective cohort study. *Chest*, 130(4):941–946, 2006.
- Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1885–1894. JMLR. org, 2017.
- Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.
- Lage, I., Ross, A., Gershman, S. J., Kim, B., and Doshi-Velez, F. Human-in-the-loop interpretability prior. In *Advances in Neural Information Processing Systems*, pp. 10159–10168, 2018.
- Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.
- Munos, R. and Moore, A. Variable resolution discretization in optimal control. *Machine learning*, 49(2-3):291–323, 2002.
- Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In

*International Conference on Machine Learning*, pp. 4881–4890, 2019.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Ribba, B., Kaloshi, G., Peyre, M., Ricard, D., Calvez, V., Tod, M., Čajavec-Bernard, B., Idbaih, A., Psimaras, D., Dainese, L., et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clinical Cancer Research*, 18(18):5071–5080, 2012.

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

Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Thomas, P. S., da Silva, B. C., Barto, A. G., Giguere, S., Brun, Y., and Brunskill, E. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

## A. Derivations for Kernel-Based FQE

### A.1. Proof of Proposition 1

**Proposition 1.** For all transitions in the dataset, the values for corresponding state-action pairs are given by

$$\hat{\mathbf{q}}'_t = \left( \sum_{t'=1}^t \gamma^{t'-1} \mathbf{M}^{t'} \right) \mathbf{r} \equiv \Phi'_t \mathbf{r} \quad (15)$$

$$\hat{\mathbf{q}}_t = \mathbf{M} \left( \sum_{t'=1}^t (\gamma \mathbf{M}')^{t'-1} \right) \mathbf{r} \equiv \Phi_t \mathbf{r}. \quad (16)$$

where  $\hat{q}'_{t,i}$  and  $\hat{q}_{t,i}$  are the estimated policy values at  $(x^{(i)}, \pi_e(x^{(i)}))$  and  $(x^{(i)}, a^{(i)})$ , respectively, for the observed transition  $i$

*Proof.* We first prove 15 by induction. We start by noting that for a given observed transition,  $i$ , averaging over all observations  $j$  such that  $\Delta_{ij}$  holds can be written as  $\frac{1}{N_i} \sum_{j:\Delta_{ij}} = \sum_j \frac{\mathbb{I}(\Delta_{ij})}{N_i} = \sum_j M_{ij}$ . Similarly, averaging over all  $j$  such that  $\Delta'_{ij}$  holds can be written as  $\sum_j M'_{ij}$ . Therefore, if  $u(x, a)$  is some function over the state-action space and  $\mathbf{u}$  is a vector containing the quantity  $u_i = u(x^{(i)}, a^{(i)})$  for every  $(x^{(i)}, a^{(i)})$ , then the nearest-neighbors estimation of  $u(x^{(i)}, \pi(x^{(i)}))$  is given by  $[\mathbf{M}'\mathbf{u}]_i$ .

Given the formulation above, for  $t = 1$ ,  $\hat{q}'_{t,i}$  estimates the reward at  $(x^{(i)}, \pi(x^{(i)}))$ , and can be written as:

$$\hat{\mathbf{q}}'_1 = \mathbf{M}' \mathbf{r}. \quad (17)$$

For  $t > 1$ , assume  $\hat{\mathbf{q}}'_{t-1} = \left( \sum_{t'=1}^{t-1} \gamma^{t'-1} \mathbf{M}^{t'} \right) \mathbf{r}$ . Then

$$\begin{aligned} \hat{\mathbf{q}}'_t &= \mathbf{M}' (\mathbf{r} + \gamma \hat{\mathbf{q}}'_{t-1}) \\ &= \mathbf{M}' \left[ \mathbf{r} + \gamma \left( \sum_{t'=1}^{t-1} \gamma^{t'-1} \mathbf{M}^{t'} \right) \mathbf{r} \right] \\ &= \left( \mathbf{M}' + \sum_{t'=1}^{t-1} \gamma^{t'} \mathbf{M}^{t'+1} \right) \mathbf{r} \\ &= \left( \sum_{t'=0}^{t-1} \gamma^{t'} \mathbf{M}^{t'+1} \right) \mathbf{r} \\ &= \left( \sum_{t'=1}^t \gamma^{t'-1} \mathbf{M}^{t'} \right) \mathbf{r} \equiv \Phi'_t \mathbf{r}, \end{aligned} \quad (18)$$

completing the proof of 15. To estimate  $\hat{\mathbf{q}}_t$ , we write  $\hat{q}_{t,i} = \frac{1}{N_i} \sum_{j:\Delta_{ij}} (r^{(j)} + \gamma \hat{q}'_{t-1,j})$  or in matrix notation.

$$\begin{aligned} \hat{\mathbf{q}}_t &= \mathbf{M} (\mathbf{r} + \gamma \hat{\mathbf{q}}'_{t-1}) \\ &= \mathbf{M} \left( \mathbf{I} + \gamma \left( \sum_{t'=1}^{t-1} \gamma^{t'-1} \mathbf{M}^{t'} \right) \right) \mathbf{r} \\ &= \mathbf{M} \left( \sum_{t'=1}^t (\gamma \mathbf{M}')^{t'-1} \right) \mathbf{r} \equiv \Phi_t \mathbf{r}. \end{aligned} \quad (19)$$

□

### A.2. Computational complexity.

Computation of a single influence value,  $\tilde{I}_{i,j}$  requires summation over all transitions  $k$  that satisfy  $\Delta_{k'j}$ . Denote the number of such neighbors by  $N_{j'}^*$ <sup>1</sup>. We expect  $N_{j'}^*$  to be small and not scale with the size of the dataset, and also  $\tilde{I}_{i,j}$  is inversely proportional to  $N_{j'}^*$ . Thus, if we only compute the influence of transitions such that  $N_{j'}^* < \frac{v_{\max}}{\delta \tilde{I}_c} \equiv N_{j',c}^*$ , where  $v_{\max}$  is the maximum possible value, we are guaranteed not to miss any transitions with influence larger than our threshold  $\tilde{I}_c$ . Since  $N_{j',c}^*$  does not scale with the size of the data, computation of a single individual influence can effectively be done in constant time. Performing influence analysis on a full dataset requires computing the influences of all transitions on all initial transitions, and therefore takes  $\mathcal{O}(N|\mathcal{D}_0^*|)$  time.

In our matrix formulation, the FQE evaluation itself is bottlenecked by computing the matrix  $\Phi$ , which includes computation of powers of  $\mathbf{M}'$ . Because  $\mathbf{M}'$  is a sparse matrix (each row  $i$  only has  $N_i$  nonzero elements), the matrix multiplication itself can be done in  $\mathcal{O}(N^2)$  rather than  $\mathcal{O}(N^3)$  time, and the entire evaluation is done in  $\mathcal{O}(N^2T)$  time. Importantly, the influence analysis analyzing all transitions has lower complexity than the OPE, and should not significantly increase the computational cost of the evaluation pipeline.

## B. Derivations Linear Least-Squares FQE

**Proposition 2.** The the linear least square solution of fitted  $Q$  evaluation is  $(\Psi^\top \Psi - \gamma \Psi^\top \Psi_p)^{-1} \Psi^\top \mathbf{r}$

*Proof.* The least-square solution of parameter vector  $\mathbf{w}$  can be found by minimizing the following square error of the Bellman equation for all  $(x, a)$  in the dataset:

$$(\hat{q}(x, a) - r(x, a) - \gamma \hat{q}(x', \pi_e(x')))^2 \quad (20)$$

Plugging in  $\hat{q}(x, a) = \boldsymbol{\psi}(x, a)^\top \mathbf{w}$ , the square error is

$$(\boldsymbol{\psi}(x, a)^\top \mathbf{w} - r(x, a) - \gamma \boldsymbol{\psi}_\pi(x')^\top \mathbf{w})^2 \quad (21)$$

<sup>1</sup>Note that  $N_{j'}^*$ , which counts all  $k$  that satisfy  $\Delta_{k'j}$ , is subtly different from the quantity  $N_{j'}$  introduced in section 5.1, which counts all  $k$  that satisfy  $\Delta_{j'k}$ .

By definition of  $\Psi$  and  $\Psi_p$ , the mean square error over the  $N$  samples is:

$$\|\Psi \mathbf{w} - \mathbf{r} - \gamma \Psi_p \mathbf{w}\|_2^2 \quad (22)$$

The least square solution is:

$$\mathbf{w} = (\Psi^\top \Psi)^{-1} \Psi^\top (\mathbf{r} + \gamma \Psi_p \mathbf{w}) \quad (23)$$

$$(\Psi^\top \Psi) \mathbf{w} = \Psi^\top (\mathbf{r} + \gamma \Psi_p \mathbf{w}) \quad (24)$$

$$(\Psi^\top \Psi - \gamma \Psi^\top \Psi_p) \mathbf{w} = \Psi^\top \mathbf{r} \quad (25)$$

$$\mathbf{w} = (\Psi^\top \Psi - \gamma \Psi^\top \Psi_p)^{-1} \Psi^\top \mathbf{r} \quad (26)$$

□

**Proposition 3.** Let  $\mathbf{C}_{-j} = (\Psi_{-j}^\top \Psi_{-j} - \gamma \Psi_{-j}^\top \Psi_{p,-j})$  and  $\mathbf{C} = (\Psi^\top \Psi - \gamma \Psi^\top \Psi_p)$ .

$$\begin{aligned} \mathbf{w}_{-j} &= (\mathbf{C}_{-j})^{-1} \left( \Psi_{-j}^\top \mathbf{r} - r^{(j)} \boldsymbol{\psi}_j \right) \\ (\mathbf{C}_{-j})^{-1} &= \mathbf{B}_j - \frac{\gamma \mathbf{B}_j \boldsymbol{\psi}_j \boldsymbol{\psi}_{\pi,j}^\top \mathbf{B}_j}{1 + \gamma \boldsymbol{\psi}_{p,j}^\top \mathbf{B}_j \boldsymbol{\psi}_j} \end{aligned}$$

where

$$\mathbf{B}_j = \mathbf{C}^{-1} + \frac{\mathbf{C}^{-1} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top \mathbf{C}^{-1}}{1 - \boldsymbol{\psi}_j^\top \mathbf{C}^{-1} \boldsymbol{\psi}_j}$$

*Proof.* By the list squares solution of FQE,  $\mathbf{w}_{-j}$  equals  $(\Psi_{-j}^\top \Psi_{-j} - \gamma \Psi_{-j}^\top \Psi_{p,-j})^{-1} \Psi_{-j}^\top \mathbf{r}_{-j}$ . Since  $\Psi_{-j}^\top \mathbf{r}_{-j} = \Psi^\top \mathbf{r} - r^{(j)} \boldsymbol{\psi}_j$ , we have that  $\mathbf{w}_{-j} = (\mathbf{C}_{-j})^{-1} (\Psi^\top \mathbf{r} - r^{(j)} \boldsymbol{\psi}_j)$ . Then

$$\mathbf{C}_{-j} = \mathbf{C} - \boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top + \gamma \boldsymbol{\psi}_j \boldsymbol{\psi}_{\pi,j}^\top, \quad (27)$$

because

$$\Psi_{-j}^\top \Psi_{-j} = \Psi^\top \Psi - \boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top \quad (28)$$

$$\Psi_{-j}^\top \Psi_{p,-j} = \Psi^\top \Psi_p - \boldsymbol{\psi}_j \boldsymbol{\psi}_{\pi,j}^\top \quad (29)$$

This indicate  $\mathbf{C}_{-j}$  equals  $\mathbf{C}$  plus two rank-1 matrices. Fortunately, we can store  $\mathbf{C}^{-1}$  when we compute  $\mathbf{w}$  and  $\hat{q}$ . The following result named Sherman–Morrison formula allow us to compute  $\mathbf{C}_{-j}^{-1}$  from  $\mathbf{C}^{-1}$  in an efficient way. For any invertible matrix  $\mathbf{A} \in \mathbb{R}^{d \times d}$  and vector  $u, v \in \mathbb{R}^d$ :

$$(\mathbf{A} + uv^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} u v^\top \mathbf{A}^{-1}}{1 + v^\top \mathbf{A}^{-1} u} \quad (30)$$

Then if we define  $\mathbf{B}_j \equiv (\mathbf{C} - \boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top)^{-1}$ , we have that

$$\mathbf{B}_j = \mathbf{C}^{-1} + \frac{\mathbf{C}^{-1} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top \mathbf{C}^{-1}}{1 - \boldsymbol{\psi}_j^\top \mathbf{C}^{-1} \boldsymbol{\psi}_j} \quad (31)$$

$$(\mathbf{C}_{-j})^{-1} = \mathbf{B}_j - \frac{\gamma \mathbf{B}_j \boldsymbol{\psi}_j \boldsymbol{\psi}_{\pi,j}^\top \mathbf{B}_j}{1 + \gamma \boldsymbol{\psi}_{p,j}^\top \mathbf{B}_j \boldsymbol{\psi}_j} \quad (32)$$

□

## C. Preprocessing and experimental details for MIMIC-III acute hypotension dataset

In this section, we describe the preprocessing we performed on the raw MIMIC-III database to convert it into a dataset amenable to modeling with RL. This preprocessing procedure was done in close consultation with the intensivist collaborator on our team.

### C.1. Cohort Selection

We use MIMIC-III v1.4 (Johnson et al., 2016), which contains information from about 60,000 intensive care unit (ICU) admissions to Beth Israel Deaconess Medical Center. We filter the initial database on the following features: admissions where data was collected using the Metavision clinical information system; admissions to a medical ICU (MICU); adults (age  $\geq 18$  years); initial ICU stays for hospital admissions with multiple ICU stays; ICU stays with a total length of stay of at least 24 hours; and ICU stays where there are 7 or more mean arterial pressure (MAP) values of 65mmHg or less, indicating probable acute hypotension. For long ICU stays, we limit to only using information captured during the initial 48 hours after admission, as our intensivist advised that care for hypotension during later periods of an ICU stay often look very different. After this filtering, we have a final cohort consisting of 1733 distinct ICU admissions. For computational convenience, we further down-sample this cohort, and use 20% (346) ICU stays to use to learn a policy, and another 20% (346) ICU stays to evaluate the policy via FQE and our proposed influence analysis.

### C.2. Clinical Variables Considered

Given our final cohort of patients admitted to the ICU, we next discuss the different clinical variables that we extract that are relevant to our task of acute hypotension management.

The two first-line treatments are intravenous (IV) fluid bolus therapy, and vasopressor therapy. We construct fluid bolus variables in the following way:

1. We filter all fluid administration events to only include NaCl 0.9%, lactated ringers, or blood transfusions (packed red blood cells, fresh frozen plasma, or platelets).
2. Since a fluid bolus should be a nontrivial amount of fluid administered over a brief period of time, we further filter to only fluid administrations with a volume of at least 250mL and over a period of 60 minutes or shorter.

Each fluid bolus has an associated volume, and a starting

time (since a bolus is given quickly / near-instantaneously, we ignore the end-time of the administration). To construct vasopressors, we first normalize vasopressor infusion rates across different drug types as follows, using the same normalization as in Komorowski et al. (2018):

1. Norepinephrine: this is our “base” drug, as it’s the most commonly administered. We will normalize all other drugs in terms of this drug. Units for vasopressor rates are in mcg per kg body weight per minute for all drugs except vasopressin.
2. Vasopressin: the original units are in units/min. We first clip any values above 0.2 units/min, and then multiply the final rates by 5.
3. Phenylephrine: we multiply the original rate by 0.45.
4. Dopamine: we multiply the original rate by 0.01.
5. Epinephrine: this drug is on the same scale as norepinephrine and is not rescaled.

As vasopressors are given as a continuous infusion, they consist of both a treatment start time and stop time, as well as potentially many times in the middle where the rates are changed. More than a single vasopressor may be administered at once, as well.

We also use 11 other clinical variables as part of the state space in our application: serum creatinine,  $\text{FiO}_2$ , lactate, urine output, ALT, AST, diastolic/systolic blood pressure, mean arterial pressure (MAP; the main blood pressure variable of interest),  $\text{PO}_2$ , and the Glasgow Coma Score (GCS).

### C.3. Selecting Action Times

Given a final cohort, clinical variables, and treatment variables, we still must determine how to discretize time and choose at which specific time points actions should be chosen. To arrive at a final set of “action” times for a specific ICU stay, we use the following heuristic-based algorithm:

1. We start by including all times a treatment is started, stopped, or modified.
2. Next, we remove consecutive treatment times if there are no MAP measurements between treatments. We do this because without at least one MAP measurement in between treatments, we would not be able to assess what effect the treatment had on blood pressure. This leaves us with a set of time points when treatments were started or modified.
3. At many time points, the clinician consciously chooses not to take an action. Unfortunately, this information is not generally recorded (although, on occasion,

may exist in clinical notes). As a proxy, we consecutively add to our existing set of “action times” any time point at which an abnormally low MAP is observed ( $< 60\text{mmHg}$ ) and there are no other “action times” within a 1 hour window either before or after. This captures the relatively fine-granularity with which a physician may choose not to treat despite some degree of hypotension.

4. Last, we add additional time points to fill in any large gaps where no “action times” exist. We do this by adding time points between existing “action times” until there are no longer any gaps greater than 4 hours between actions. This makes some clinical sense, as patients in the ICU are being monitored relatively closely, but if they are more stable, their treatment decisions will be made on a coarser time scale.

Now that we have a set of action times for each trajectory, we can count up the total number of transitions in our training and evaluation datasets (both of which consist of 346 trajectories). The training trajectories contain a total of 6777 transitions, while there are 6863 total transitions in the evaluation data. Trajectories vary in length from a minimum of 7 transitions to a maximum of 49, with 16, 18, and 23 transitions comprising the 25%, 50%, and 75% quantiles, respectively.

### C.4. Action Space Construction

Given treatment timings, doses, and manually identified “action times” at which we want to assess what type of clinical decision was made, we can now construct our action space. We choose to operate in a discrete action space, which means we need to decide how to bin each of the continuous-valued treatment amounts.

Binning of IV fluids is more natural and easier, as fluid boluses are generally given in discrete amounts. The most common bolus sizes are 500mL and 1000mL, so we bin fluid bolus volumes into the following 4 bins, which correspond to “none”/“low”/“medium”/“high” (in mL):  $\{0, [250, 500), [500, 1000), [1000, \infty)\}$ , although in practice very few boluses of more than 2L are ever given. Given this binning scheme, we can simply add up the total amount of fluids administered during any adjacent action times to determine which discrete fluid amount we should code the action as.

Binning of vasopressors is slightly more complex. These drugs are dosed at a specific rate, and there may be many rate changes made between action times, or sometimes there are several vasopressors being given at once. We chose to first add up the cumulative amount of (normalized) vasopressor drug administered between action times, and then normalize this amount by the size of the time

window between action times to account for the irregular spacing. Finally, we also bin vasopressors into 4 discrete bins corresponding to “none”/“low”/“medium”/“high” amounts:  $\{0, (0, 8.1), [8.1, 21.58), [21.58, \infty)\}$ . The relevant units here are total mcg of drug given each hour, per kg body weight. Since the distribution of values for vasopressors is not as naturally discrete, we chose our bin sizes using the 33.3% and 66.7% quantiles of dose amounts.

In the end, we have an action space with 16 possible discrete actions, considering all combinations of each of the 4 vasopressor amounts and fluid bolus amounts.

### C.5. State Construction

Given a patient cohort, decision/action times, and discrete actions, we are now ready to construct a state space. For simplicity in this initial work, we first start with the 11 clinical time series variables previously listed. If a variable is never measured, we use the population median as a placeholder. If a variable has been measured before, we use the most recent measurement. The sole exception to this is the 3 blood pressure variables. For the blood pressures, we instead use the minimum (or worst) value observed since the last action.

We add to these a number of indicator variables that denote whether a particular variable was recently measured or not. Due to the strongly missing-not-at-random nature of clinical time series, there is often considerable signal in knowing that certain types of measurements were recently taken, irrespective of the measurement values (Agniel et al., 2018). We choose to construct indicator variables denoting whether or not a urine output was taken since the last action time, and whether a GCS was recorded since the last action. We also include state features denoting whether the following labs/vitals were *ever* ordered: creatinine, FiO<sub>2</sub>, lactate, ALT, AST, PO<sub>2</sub>. We do not include these indicators for all 11 clinical variables, as most of the vitals are recorded at least once an hour, and sometimes even more frequently. In total, 8 indicators comprise part of our state space.

Last, we include 10 additional variables that summarize past treatments administered, if any. We first include 6 indicator variables (3 for each treatment type) denoting which dose of fluid and vasopressor, if any, was chosen at the last action time. Last, for each treatment type we include two final features that summarize past actual amounts of treatments administered (the total amount of this treatment administered up until the current time, and the total amount of this treatment administered within the last 8 actions).

In total, our final state space has 29 dimensions. In future work we plan to explore richer state representations.

### C.6. Reward Function Construction

In this preliminary work, we use a simple reward that is a piecewise linear function of the MAP in the next state. In particular, the reward takes on a value of  $-1$  at 40mmHg, the lowest attainable MAP in the data. It increases linearly to  $-0.15$  at 55mmHg, linearly from there to  $-0.05$  at 60mmHg, and achieves a maximum value of  $0$  at 65mmHg, a commonly used target for blood pressure in the ICU (Asfar et al., 2014). However, if a patient has a urine output of 30mL/hour or higher, then any MAP values of 55mmHg or higher are reset to  $0$ . This attempts to mimic the fact that a clinician will not be too concerned if a patient is slightly hypotensive but otherwise stable, since a modest urine output indicates that the modest hypotension is not a real problem.

### C.7. Choice of Kernel Function

In order to use kernel-based FQE, we need to define a kernel that defines similarity between states. In consultation with our intensivist collaborator, we chose a simple weighted Euclidean distance, where each state variable receives a different weight based on its estimated importance to the clinical problem. We show all weights in Table 1.

Since technically we need a kernel over both all possible states and actions for FQE and influence analysis, we augment our kernel with extremely large weights so that effectively the kernel only compares pairs  $(s, a)$  and  $(s', a')$  for  $a = a'$ . Other choices should be made for continuous action spaces.

### C.8. Hyperparameters

We use the training set of 6777 trajectories to learn a policy to then evaluate using FQE and influence analysis. In particular, we learn a deterministic policy by taking the most common action within the 50 nearest neighbors of a given state, with respect to the kernel in Table 1. We use a discount of  $\gamma = 1$  so that all time steps are treated equally, and use a neighborhood radius of 7 for finding nearest neighbors in FQE. Lastly, for the influence analysis, we use a threshold of 0.05, or 5%, so that transitions which will affect the FQE value estimate by more than 5% are flagged for expert review.

## D. Additional Results from MIMIC-III acute hypotension dataset

In the main body of the paper, we showed two qualitative results figures showing 2 of the 6 highly influential transitions flagged by influence analysis. In this section, we show the remaining 4 influential transitions.

Table 1. Weights for each state variable in our kernel function.

State Variable	Kernel Weight
Creatinine	3
FiO2	15
Lactate	10
Urine Output	15
Urine Output since last action?	5
ALT	5
AST	5
Diastolic BP	5
MAP	15
PO2	3
Systolic BP	5
GCS	15
GCS since last action?	5
Creatinine ever taken?	3
FiO2 ever taken?	15
Lactate ever taken?	10
ALT ever taken?	5
AST ever taken?	5
PO2 ever taken?	3
Low vasopressor done last time?	15
Medium vasopressor done last time?	15
High vasopressor done last time?	15
Low fluid done last time?	15
Medium fluid done last time?	15
High fluid done last time?	15
Total vasopressors so far	15
Total fluids so far	15
Total vasopressors last 8 actions	15
Total fluids last 8 actions	15

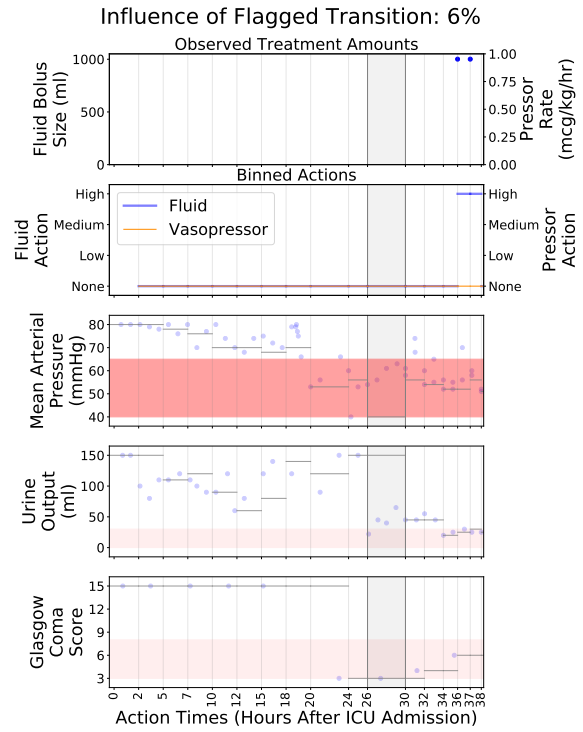


Figure 5. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate.

Interpretable Off-Policy Evaluation by Highlighting Influential Transitions

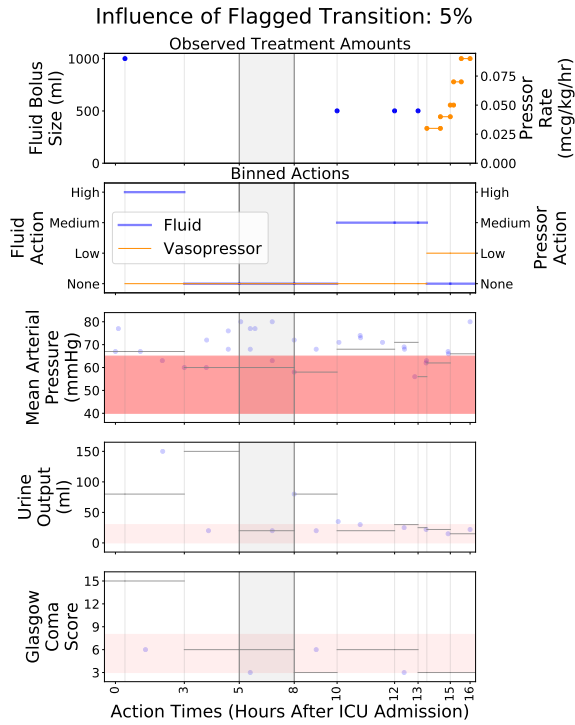


Figure 6. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate. Note that this transition is from the same trajectory as the influential transition highlighted in Figure 7

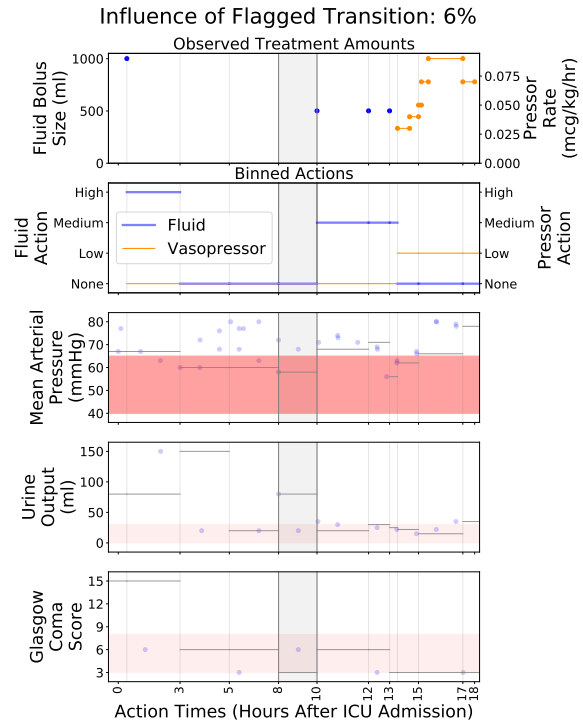


Figure 7. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate. Note that this transition is from the same trajectory as the influential transition highlighted in Figure 6



## Interpretable Off-Policy Evaluation by Highlighting Influential Transitions

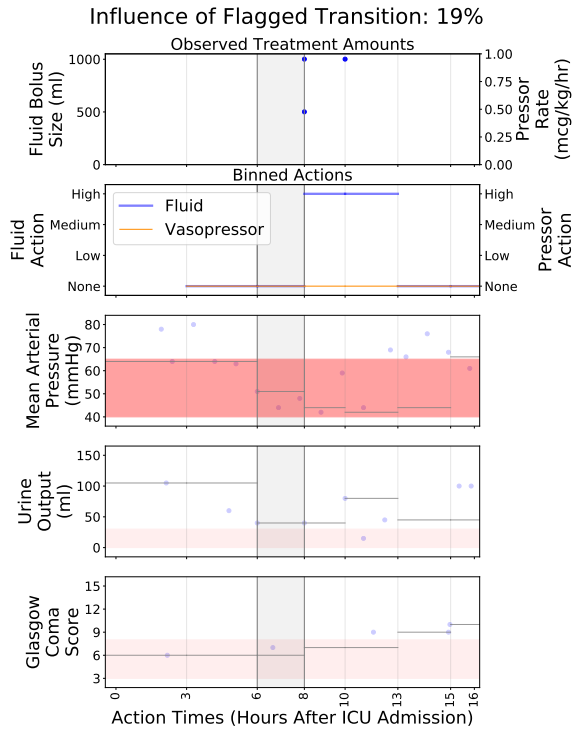


Figure 8. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate.