A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units

Niranjani Prasad
Princeton University

Li-Fang Cheng
Princeton University

Corey Chivers
Penn Medicine

Michael Draugelis
Penn Medicine

Barbara E. Engelhardt
Princeton University

Abstract

The management of invasive mechanical ventilation, and the regulation of sedation and analgesia during ventilation, constitutes a major part of the care of patients admitted to intensive care units. Both prolonged dependence on mechanical ventilation and premature extubation are associated with increased risk of complications and higher hospital costs, but clinical opinion on the best protocol for weaning patients off of a ventilator varies. This work aims to develop a decision support tool that uses available patient information to predict time-to-extubation readiness and to recommend a personalized regime of sedation dosage and ventilator support. To this end, we use off-policy reinforcement learning algorithms to determine the best action at a given patient state from sub-optimal historical ICU data. We compare treatment policies from fitted Q-iteration with extremely randomized trees and with feedforward neural networks, and demonstrate that the policies learnt show promise in recommending weaning protocols with improved outcomes, in terms of minimizing rates of reintubation and regulating physiological stability.

1 Introduction

Mechanical ventilation is one of the most widely used interventions in admissions to the intensive care unit (ICU): around 40% of patients in the ICU are supported on invasive mechanical ventilation at any given hour, accounting for 12% of total hospital costs in the United States (Ambrosino and Gabbrielli [2010], Wunsch et al. [2013]). These are typically patients with acute respiratory failure or compromised lung function caused by some underlying condition such as pneumonia, sepsis, or heart disease, or cases in which breathing support is necessitated by neurological disorders, impaired consciousness, or weakness following major surgery. As advances in healthcare enable more patients to survive critical illness or surgery, the need for mechanical ventilation during recovery has risen.

Closely coupled with ventilation in the care of these patients is sedation and analgesia, which are crucial to maintaining physiological stability and controlling pain levels of patients while intubated. The underlying condition of the patient, as well as factors such as obesity or genetic variations, can have a significant effect on the pharmacology of drugs, and cause high inter-patient variability in response to a given sedative (Patel and Kress [2012]), lending motivation to a personalized approach to sedation strategies.

Weaning refers to the process of liberating patients from mechanical ventilation. The primary diagnostic tests for determining whether a patient is ready to be extubated involve screening for resolution of the underlying disease, haemodynamic stability, assessment of current ventilator settings and level of consciousness, and finally a series of spontaneous breathing trials (SBTs). Prolonged ventilation—and corresponding over-sedation—is associated with post-extubation delirium, drug dependence, ventilator-induced pneumonia, and higher patient mortality rates (Hughes et al. [2012]), in addition to inflating costs and straining hospital resources. Physicians are often conservative in recognizing patient suitability for extubation, however, as failed breathing trials or premature extubations that necessitate reintubation within 48-72 hours can cause severe patient discomfort and result in even longer ICU stays (Krinsley et al. [2012]). Efficient weaning of sedation and ventilation is therefore a priority both for improving patient outcomes and reducing costs, but a lack of comprehensive evidence and the variability in outcomes between individuals and subpopulations means there is little agreement in clinical literature on the best weaning protocol (Conti et al. [2014], Goldstone [2002]).

In this work, we aim to develop a decision support tool that leverages available patient information in the data-rich ICU setting to alert clinicians when a patient is ready for initiation of weaning, and to recommend a personalized treatment protocol. We explore the use of off-policy re-

inforcement learning algorithms, namely fitted Q-iteration (FQI) with different regressors, to determine the optimal treatment at each patient state from sub-optimal historical patient treatment profiles. The setting fits naturally into the framework of reinforcement learning as it is fundamentally a sequential decision making problem rather than purely a prediction task: we wish to choose the best possible action at each time—in terms of sedation drug and dosage, ventilator settings, initiation of a spontaneous breathing trial, or extubation—while capturing the stochasticity of the underlying process, the delayed effects of actions, and the uncertainty in state transitions and outcomes.

The problem poses a number of key challenges: there are a multitude of factors that can potentially influence patient readiness for extubation, including some not directly observed in ICU chart data, such as a patient's inability to protect their airway due to muscle weakness. The data that is recorded is often sparse and noisy. In addition, there is potentially an extremely large space of possible sedatives and ventilator settings that can be leveraged during weaning. We are also posed with the problem of interval censoring, as in other intervention data: given past treatment and vitals trajectories, observing a successful extubation at time t provides us only with an upper bound on the true time to extubation readiness, $t_e \leq t$; on the other hand, if a breathing trial was unsuccessful, there is uncertainty how premature the intervention was. This presents difficulties both when learning the policy and in evaluating policies.

The rest of the paper is organized as follows: Section 2 explores recent efforts in the use of reinforcement learning in clinical settings. In Section 3, we describe the data and methods used here, and Section 4 presents the results. Finally, conclusions and possible directions for further work are discussed in Section 5.

2 Related Work

The widespread adoption of electronic health records (EHRs) paved the way for a data-driven approach to healthcare, and recent years have seen a number of efforts towards personalized, dynamic treatment regimes. Reinforcement learning in particular has been explored in various settings, from determining the sequence of drugs to be administered in HIV therapy or cancer treatment, to management of anaemia in haemodialysis patients, and insulin regulation in diabetics. These efforts are typically based on estimating the *value*, in terms of clinical outcomes, of different treatment decisions given the state of the patient.

For example, Ernst et al. [2006] applied fitted Q-iteration with a tree-based ensemble method to learn the optimal HIV treatment in the form of structured treatment interruption strategies, in which patients are cycled on and off drug therapy. The observed reward here is defined in terms of the equilibrium point between healthy and unhealthy blood cells in the patient. Zhao et al. [2011] used Q-learning to learn optimal individualized treatment regimens for nonsmall cell lung cancer. The objective is to choose the optimal first and second lines of therapy and the optimal initiation time for the second line treatment such that the overall survival time is maximized. The Q-function with time-indexed parameters is approximated using a modification of support vector regression (SVR) that explicitly handles right-censored data. In this setting, right-censoring arises in measuring the time of death from start of therapy: given that a patient is still alive at the time of the last follow-up, we merely have a lower bound on the exact survival time.

Escandell-Montero et al. [2014] compared the performance of both Q-learning and fitted Q-iteration with current clinical protocol for informing the delivery of erythropoeisis-stimulating agents (ESAs) for treating anaemia. The drug administration strategy is modeled as a Markov decision process (MDP), with the state space expressed by current values and change in haemoglobin levels, the most recent ESA dosage, and the patient subpopulation. The action space is a set of four discretized ESA levels, and the reward function is designed to maintain haemoglobin levels within a healthy range while avoiding abrupt changes.

On the problem of administering anaesthesia in an ICU setting, Moore et al. [2004] applied Q-learning with eligibility traces to the administration of intravenous propofol, modeling patient dynamics according to an established pharmacokinetic model, with the aim of maintaining some level of sedation or consciousness. Padmanabhan et al. [2014] also used Q-learning, for the regulation of both sedation level and arterial pressure as an indicator of physiological stability, using propofol infusion rate. All of the aforementioned work rely on model-based approaches to reinforcement learning, and develop treatment policies on simulated patient data. More recently however, Nemati et al. [2016] consider the problem of heparin dosing to maintain blood coagulation levels within some well-defined therapeutic range, modeling the task as a partially observable MDP, using a dynamic Bayesian network trained on real ICU data, and learning a dosing policy with neural fitted Q-iteration (NFQ).

There exists some literature on machine learning methods for the problem of ventilator weaning: Mueller et al. [2013] and Kuo et al. [2015] look at prediction of weaning outcomes using supervised learning methods, and suggest that classifiers based on neural networks, logistic regression, or naive Bayes, trained on patient ventilator and blood gas data, show promise in predicting successful extubation. Gao et al. [2017] develops association rule networks for naive Bayes classifiers, to analyze the discriminative power of different feature categories toward each decision outcome class, in order to help inform clinical decision making. Our paper is novel in its use of reinforcement learning methods to directly tackle policy recommendation for

ventilation weaning. Specifically, we incorporate a larger number of possible predictors of weaning readiness, in a 32-dimensional patient state representation, compared with previous works which typically limit features for classification to at most a couple of key vital signs. Moreover, we make use of current clinical protocols to inform the design and tuning of a reward function.

3 Methods

3.1 Critical Care Data

We use the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC III) database (Johnson et al. [2016]), a freely available source of de-identified critical care data for 53,423 adult admissions and 7,870 neonates. The data includes patient demographics, time-stamped measurements from bedside monitoring of vitals, administration of fluids and medications, results of laboratory tests, observations and notes charted by care providers, as well as diagnoses, procedures and prescriptions for billing.

We extract from this database a set of 8,860 admissions from 8,182 unique adult patients undergoing invasive ventilation. In order to train and test our weaning policy, we filter further to include only those admissions in which the patient was kept under ventilator support for more than 24 hours. This allows us to exclude the majority of episodes of routine ventilation following surgery, which are at minimal risk of adverse extubation outcomes. We also filter out admissions in which the patient in not successfully discharged from the hospital by the end of the admission, as in cases where the patient expires in the ICU, failure to discharge is largely due to factors beyond the scope of ventilator weaning, and again, a more informed weaning policy is unlikely to have a significant influence on outcomes. *Failure* in our problem setting is instead defined as prolonged ventilation, administration of unsuccessful spontaneous breathing trials, or reintubation within the same admission—all of which are associated with adverse outcomes for the patient. A typical patient timeline is illustrated in Figure 1.

Preliminary guidelines for the weaning protocol, in terms of the desired ranges of physiological parameters (heart rate, respiratory rate, and arterial pH) as well as criteria at time of extubation for the inspired O_2 fraction (FiO_2), oxygenation pulse oxymetry (SpO_2), and positive end-expiratory pressure (PEEP) set, were obtained from clinicians at the Hospital of University of Pennsylvania, HUP (Table 1). These are used in shaping rewards in our MDP to facilitate learning of the optimal policy.

Physiological Stability	**Oxygenation Criteria**
Respiratory Rate ≤ 30	PEEP (cm H_2O) ≤ 8
Heart Rate ≤ 130	SpO_2 (%) ≥ 88
Arterial pH ≥ 7.3	Inspired O_2 (%) ≤ 50

Table 1: Current extubation guidelines at HUP.

3.1.1 Preprocessing using Gaussian Processes

Measurements of vitals and lab results in the ICU data can be irregular, sparse, and error-prone. Non-invasive measurements such as heart rate or respiratory rate are taken several times an hour, while tests for arterial pH or oxygen pressure, which involve more distress to the patient, may only be administered every few hours as needed. This wide discrepancy in measurement frequency is typically handled by resampling with means in hourly intervals (when we have multiple measurements within an hour), and using sample-and-hold interpolation to impute subsequent missing values. However, patient state—and therefore the need to update management of sedation or ventilation—can change within the space of an hour, and naive methods for interpolation are unlikely to provide the necessary accuracy at higher temporal resolutions. We therefore explore methods for the imputation of patient state that can enable more precise policy estimation.

One commonly used approach to resolve missing and irregularly sampled time series data is Gaussian processes (GPs, [Stegle et al., 2008, Dürichen et al., 2015, Ghassemi et al., 2015]).Denoting the observations of the vital signs by $\mathbf{v}$ and the measurement time $\mathbf{t}$, we model

$$\mathbf{v} = f(\mathbf{t}) + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ vector represents i.i.d Gaussian noise, and $f(\mathbf{t})$ is the latent noise-free function we would like to estimate. We put a GP prior on the latent function $f(\mathbf{t})$:

$$f(\mathbf{t}) \sim \mathcal{GP}(m(\mathbf{t}), \kappa(\mathbf{t}, \mathbf{t}')),$$

where $m(\mathbf{t})$ is the *mean function* and $\kappa(\mathbf{t}, \mathbf{t}')$ is the *covariance function* or *kernel*, which shapes the temporal properties of $f(\mathbf{t})$. In this work, we use a multi-output GP to account for temporal correlations between physiological signals during interpolation. We adapt the framework in Cheng et al. [2017] to impute the physiological signals jointly by estimating covariance structures between them, excluding the sparse prior settings. We set $m(\mathbf{t}) = \mathbf{0}$ without loss of generality (Rasmussen and Williams [2006]), and $\kappa(\mathbf{t}, \mathbf{t}')$ as the kernel in the linear model of coregionalization with the spectral kernel as the basis kernel, allowing us to model both smooth correlations in time and periodic variations of these vital signs and lab results. The full joint kernel for each patient i is defined as:

$$\kappa_i(\mathbf{t}_i, \mathbf{t}_i') = \sum_{q=1}^{Q} \boldsymbol{B}_q \otimes \kappa_q(\mathbf{t}_{i,*}, \mathbf{t}_{i,*}'),$$

where $\mathbf{t}_{i,*}$ represents the time vector of each vital sign. Note that this is a simplified representation based on the

Figure 1: Example ventilated ICU patient. Vitals are measured at a range of sampling intervals. Ventilation times are marked, and multiple administered sedatives (both as continuous IV drips and discrete boli) are shown.

assumption that we have the same input time vector for each signal, which does not hold in our irregularly sampled data. In practice we have to compute each sub-block $\kappa_q(\mathbf{t}_{i,d}, \mathbf{t}'_{i,d'})$ given any pair of input time $\mathbf{t}_{i,d}$ and $\mathbf{t}'_{i,d'}$ from two signals, indexing by d and d'. We use Q to denote the number of mixture kernels, and $\boldsymbol{B}_q$ to encode the scale covariance between any pair of signals, written as

$$\boldsymbol{B}_q = \begin{bmatrix} b_{q,(1,1)} & b_{q,(1,2)} & \cdots & b_{q,(1,D)} \\ b_{q,(1,1)} & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ b_{q,(D,1)} & b_{q,(D,2)} & \cdots & b_{q,(D,D)} \end{bmatrix} \in \mathbb{R}^{D \times D}.$$

The basis kernel is parameterized as

$$\kappa_q(\mathbf{t}, \mathbf{t}') = \exp\left(-2\pi^2\tau^2 v_q\right) \cos\left(2\pi\tau\mu_q\right),$$

$$\tau = |\mathbf{t} - \mathbf{t}'|.$$

We set $Q = 2$ and $R = 5$ for modeling 12 selected physiological signals ($D = 12$) jointly. For each patient, one structured GP kernel is estimated using the implementation in Cheng et al. [2017]. We then impute the time series with the estimated posterior mean given all the observations across all chosen physiological signals for that patient. In choosing the 12 signals, we exclude vitals that take discrete values, such as ventilator mode or the RASS sedation scale; for these, we simply resample with means and apply sample-and-hold interpolation. After preprocessing, we obtain complete data for each patient, at a temporal resolution of 10 minutes, from admission time to discharge time (Figure 2).

3.2 MDP Formulation

A Markov decision process is defined by

(i) A finite **state space** $\mathcal{S}$ such that at each time t, the environment (here, the patient) is in state $s_t \in \mathcal{S}$.

(ii) An **action space** $\mathcal{A}$: at each time t, the agent takes action $a_t \in \mathcal{A}$, which influences the next state, s_{t+1}.

(iii) A **transition function** $P(s_{t+1}|s_t, a_t)$, the probability of the next state given the current state and action, which defines the (unknown) dynamics of the system.

(iv) A **reward function** $r(s_t, a_t) \in \mathbb{R}$, the observed feedback following a transition at each time step t.

The goal of the reinforcement learning agent is to learn a **policy**, i.e. a mapping $\pi(s) \to a$ from states to actions, that maximizes the expected accumulated reward

$$R^\pi(s_t) = \lim_{T\to\infty} \mathbb{E}_{s_{t+1}|s_{t,\pi(s_t)}} \sum_{t=1}^{T} \gamma^t r(s_t, a_t)$$

over time horizon T. The discount factor, γ, determines the relative weight of immediate and long-term rewards.

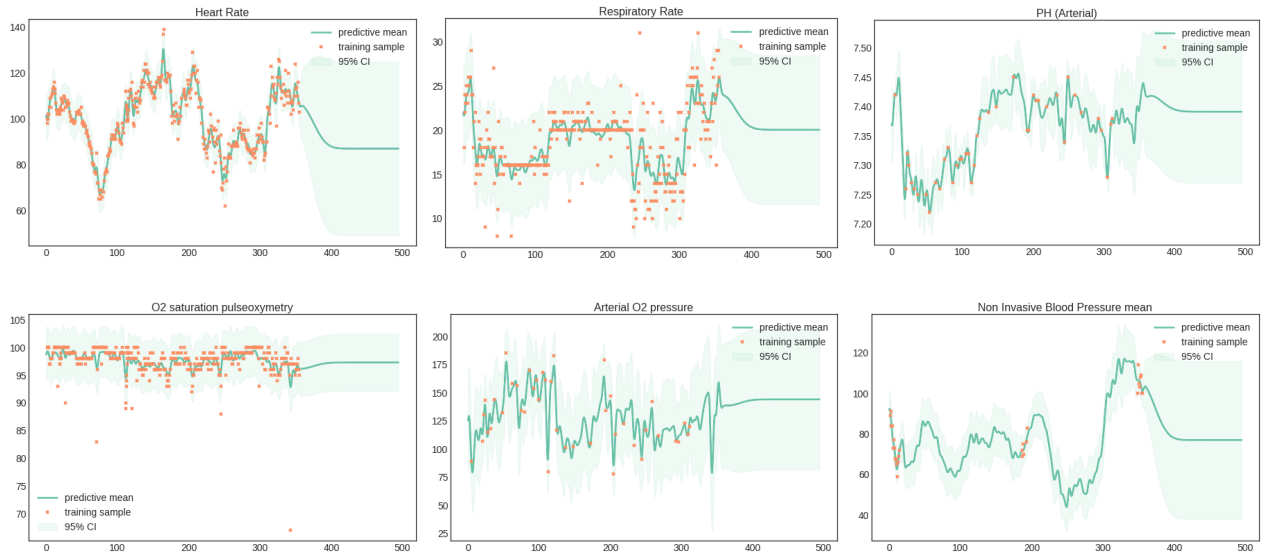


Figure 2: Example trajectories of six vital signs for a single admission, following imputation using Gaussian processes. Twelve vital signs are jointly modeled by the GP.

Patient response to sedation and readiness for extubation can depend on a number of different factors, from demographic characteristics, pre-existing conditions, and comorbidities to specific time-varying vital signs, and there is considerable variability in clinical opinion on the extent of the influence of different factors. Here, in defining each patient state within an MDP, we look to incorporate as many reliable and frequently monitored features as possible, and allow the algorithm to determine the relevant features. The state at each time t is a 32-dimensional feature vector that includes fixed demographic information (patient *age, weight, gender, admit type, ethnicity*), as well as relevant physiological measurements, ventilator settings, level of consciousness (given by the Richmond Agitation Sedation Scale, or RASS), current dosages of different sedatives, time into ventilation, and the number of intubations so far in the admission. For simplicity, categorical variables *admit type* and *ethnicity* are binarized as emergency/non-emergency and white/non-white, respectively.

In designing the action space, we develop an approximate mapping of six commonly used sedatives into a single dosage scale, and choose to discretize this scale to four different levels of sedation. The action $a_t \in \mathcal{A}$ at each time step is chosen from a finite two-dimensional set of eight actions, where $a_t[0] \in \{0, 1\}$ indicates having the patient off or on the ventilator, respectively, and $a_t[1] \in \{0, 1, 2, 3\}$ corresponds to the level of sedation to be administered over the next 10-minute interval:

$$\mathcal{A} = \left\{ \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0\\3\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}1\\3\end{bmatrix} \right\}$$

Finally, we associate a reward signal r_{t+1} with each state transition—defined by the tuple $\langle s_t, a_t, s_{t+1} \rangle$—to encompass (i) time into ventilation, (ii) physiological stability, i.e. whether vitals are steady and within expected ranges, (iii) failed SBTs or reintubation. The reward at each timestep is defined by a combination of sigmoid, piecewise-linear, and threshold functions that reward closely regulated vitals and successful extubation while penalizing adverse events:

$r_{t+1} = r^{vitals}_{t+1} + r^{vent\ off}_{t+1} + r^{vent\ on}_{t+1}$, where

$$r^{vitals}_{t+1} = C_1 \sum_v \left[\frac{1}{1+e^{-(v_t - v_{min})}} - \frac{1}{1+e^{-(v_t - v_{max})}} + \frac{1}{2} \right] - C_2 \left[\max\left(0, \frac{|v_{t+1} - v_t|}{v_t} - 0.2\right) \right],$$

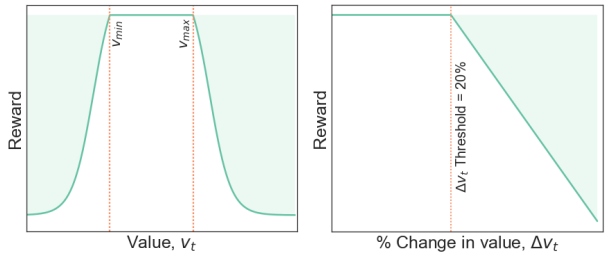
$$r^{vent\ off}_{t+1} = \mathbb{1}_{[s_{t+1}(\text{vent on})=0]} \left[C_3 \cdot \mathbb{1}_{[s_t(\text{vent on})=1]} + C_4 \cdot \mathbb{1}_{[s_t(\text{vent on})=0]} - C_5 \sum_{v^{ext}} \mathbb{1}_{[v^{ext}_t > v^{ext}_{max} \,||\, v^{ext}_t < v^{ext}_{min}]} \right],$$

$$r^{vent\ on}_{t+1} = \mathbb{1}_{[s_{t+1}(\text{vent on})=1]} \left[C_6 \cdot \mathbb{1}_{[s_t(\text{vent on})=1]} - C_7 \cdot \mathbb{1}_{[s_t(\text{vent on})=0]} \right].$$

Here, values v_t are the measurements of those vitals v (included in the state representation s_t) believed to be indicative of physiological stability at time t, with desired ranges $[v_{min}, v_{max}]$. The penalty for exceeding these ranges at each time step is given by a truncated sigmoid function (Figure 3a). The system also receives negative feedback when consecutive measurements see a sharp change (Figure 3b).

Vital signs v^{ext}_t comprise the subset of parameters directly associated with readiness for extubation (FiO_2, SpO_2, and PEEP set) with weaning criteria defined by the ranges $[v^{ext}_{min}, v^{ext}_{max}]$. A fixed penalty is applied when these criteria are not met during extubation. The system also accumulates negative rewards for each additional hour spent on the ventilator, and a large positive reward at the time of suc-

cessful extubation. Constants C_1 to C_7 determine the relative importance of these reward signals.



(a) Exceeding threshold values (b) High fluctuation in values

Figure 3: Shape of reward from vitals, $r_t^{vitals}(v_t)$

3.3 Learning the Optimal Policy

The majority of reinforcement learning algorithms are based on estimation of the Q-function, that is, the expected value of state-action pairs $Q^\pi(s,a) : S \times A \to \mathbb{R}$, to determine the optimal policy π. Of these, the most widely used is Q-learning, an off-policy reinforcement learning algorithm in which we start with an initial state and arbitrary approximation of the Q-function, and update this estimate using the reward from the next transition using the Bellman recursion for Q-values:

$$\hat{Q}(s_t,a_t) = \hat{Q}(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_{a\in\mathcal{A}} \hat{Q}(s_{t+1},a) - \hat{Q}(s_t,a_t))$$

where the learning rate α gives the relative weight of the current and previous estimate, and γ is the discount factor.

Fitted Q-iteration (FQI) is a form of off-policy *batch-mode* reinforcement learning that uses a set of one-step transition tuples:

$$\mathcal{F} = \{(\langle s_t^n, a_t^n, s_{t+1}^n\rangle, r_{t+1}^n), n = 1, ..., |\mathcal{F}|\}$$

to learn a sequence of function approximators $\hat{Q}_1, \hat{Q}_2 ... \hat{Q}_K$ of the value of state-action pairs, by iteratively solving supervised learning problems. Both FQI and Q-learning belong to the class of model-free reinforcement learning methods, which assumes no knowledge of the dynamics of the system. In the case of FQI, there are also no assumptions made on the ordering of tuples; these could correspond to a sequence of transitions from a single admission, or randomly ordered transitions from multiple histories. FQI is therefore more data-efficient, with the full set of samples used by the algorithm at every iteration, and hence typically converges much faster than Q-learning.

The training set at the k^{th} supervised learning problem is given by $\mathcal{TS} = \{(\langle s_t^n, a_t^n\rangle, \hat{Q}_k(s_t^n, a_t^n)), n = 1, ..., |\mathcal{F}|\}$. As before, the Q-function is updated at each iteration according to the Bellman equation:

$$\hat{Q}_k(s_t,a_t) \leftarrow r_{t+1} + \gamma \max_{a\in\mathcal{A}} \hat{Q}_{k-1}(s_{t+1},a)$$

where $\hat{Q}_1(s_t,a_t) = r_{t+1}$. The approximation of the optimal policy after K iterations is then given by:

$$\hat{\pi}^*(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}_K(s,a)$$

FQI guarantees convergence for many commonly used regressors, including kernel-based methods (Ormoneit and Sen [2002]) and decision trees. In particular, extremely randomized trees (Extra-Trees: Geurts et al. [2006], Ernst et al. [2005]), a tree-based ensemble method that extends on random forests by introducing randomness in the thresholds chosen at each split, has been applied in the past to learning large or continuous Q-functions in clinical settings (Ernst et al. [2006], Escandell-Montero et al. [2014]).

Neural Fitted-Q (NFQ, Riedmiller [2005]) on the other hand, looks to leverage the representational power of neural networks as regressors to fitted Q-iteration. Nemati et al. [2016] use NFQ to learn optimal heparin dosages, mapping the patient hidden state to expected return. Neural networks hold an advantage over tree-based methods in iterative settings in that it is possible to simply update network weights at each iteration, rather than rebuilding the trees entirely.

Algorithm 1 Fitted Q-iteration with sampling

Input:
One-step transitions $\mathcal{F} = \{\langle s_t^n, a_t^n, s_{t+1}^n\rangle, r_{t+1}^n\}_{n=1:|\mathcal{F}|}$;
Regression parameters θ;
Action space $\mathcal{A}$; subset size N
Initialize $Q_0(s_t,a_t) = 0 \quad \forall s_t \in \mathcal{F},\ a_t \in \mathcal{A}$
for *iteration* $k = 1 \to K$ **do**
  $subset_N \sim \mathcal{F}$
  $S \leftarrow []$
  for $i \in subset_N$ **do**
    $Q_k(s_i,a_i) \leftarrow r_{i+1} + \gamma \max_{a'\in\mathcal{A}} (predict(\langle s_{i+1}, a'\rangle, \theta))$
    $S \leftarrow append(S, \langle (s_i,a_i),\ Q(s_i,a_i)\rangle)$
  end
  $\theta \leftarrow regress(S)$
end
Result: θ
$\pi \leftarrow classify(\langle s_t^n, a_t^n\rangle)$

4 Experimental Results

After extracting relevant ventilation episodes from ICU admissions in the MIMIC III database (Section 3.1), and splitting these into training and test data, we obtain a total of 1,800 distinct admissions in our training set and 664 admissions in our test set. We interpolate time-varying vitals measurements using Gaussian processes or sample-and-hold interpolation, sampling at 10-minute intervals. This yields of the order of 1.5 million one-step transitions in the training set and 0.5 million in the test set respectively, where each transition is a 32-dimensional representation of patient state.

As a baseline, we applied Q-learning to the training data to learn the mapping of continuous states to Q-values, with function approximation using a three-layer feedforward neural network. The network is trained using *Adam*, an efficient stochastic gradient-based optimizer (Kingma and Ba [2014]), and l_2 regularization of weights. Each patient admission k is treated as a distinct episode, with on the order of thousands of state transitions in each; the network weights are incrementally updated following each transition. Studying the change between successive episodes in the predicted Q-values for all state-action pairs in the training set (Figure 4), it is unclear whether the algorithm succeeds in converging within the 1,800 training episodes.

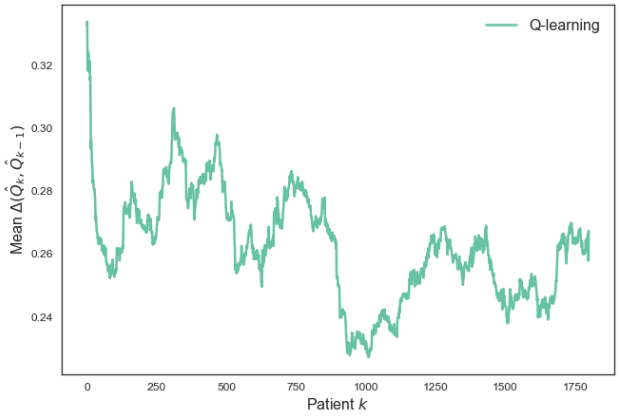


Figure 4: Convergence of $\hat{Q}(s, a)$ using Q-learning.

We then explored the use of FQI to learn our Q-function, first running with an Extra-Trees for function approximation. In our implementation, each iteration of FQI is performed on a random subset of 10% of all transitions in the training set, as described in Algorithm 1, such that on average, each sample is seen in a tenth of all iterations. Though sampling increases the total number of iterations required for convergence, it yields significant speed-ups in building trees at each iteration, and hence in total training time. The ensemble regressor learns 50 trees, with regularization in the form of a minimum leaf node size of 20 samples. We present here results with FQI performed for a fixed number of 100 iterations, though it is possible to use a convergence criterion of the form $\Delta(Q_k, Q_{k-1}) \leq \varepsilon$ for early stopping, to speed up training further.

For comparison, we used he same methods to run FQI with neural networks (NFQ) in place of tree-based regression: we train a feedforward network with architecture and techniques identical to those applied in our function approximation for Q-learning. Convergence of the estimated Q-function for both regressors is measured by the mean change in the estimate $\hat{Q}$ for transitions in the training set (Figure 5) which shows that the algorithm takes roughly 60 iterations to converge in both cases. However, NFQ yields approximately a four-fold gain in runtime speed, as expected, since with neural networks we can simply update weights rather than retraining fully at each iteration.

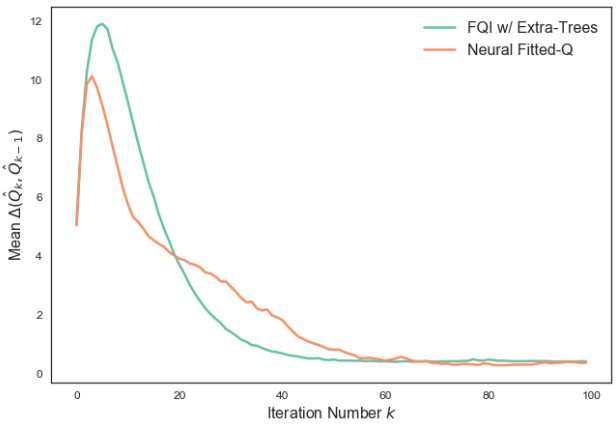


Figure 5: Convergence of estimated Q using FQI, given by the mean change in $\hat{Q}(s, a)$ over successive iterations.

The estimated Q-functions from FQI with Extra-Trees (FQIT) and from NFQ are then used to evaluate the optimal action, i.e. that which maximizes the value of the state-action pair, for each state in the training set. We can then train policy functions $\pi(s)$ mapping a given patient state to the corresponding optimal action $a \in \mathcal{A}$. To allow for clinical interpretation of the final policy, we choose to train an Extra-Trees classifier with an ensemble of 100 trees to represent the policy function.

The relative importance assigned to the top 24 features in the state space for the policy trees learnt, when training on optimal actions from both FQIT and NFQ, show that the five vitals ranking highest in importance across the two policies are arterial O_2 pressure, arterial pH, FiO_2, O_2 flow and PEEP set (Figure 6). These are as expected—Arterial pH, FiO_2, and PEEP all feature in our preliminary HUP guidelines for extubation criteria, and there is considerable literature suggesting blood gases are an important indicator of readiness for weaning (Hoo [2012]). On the other hand, oxygen saturation pulse oxymetry (SpO_2) which is also included in HUP's current extubation criteria, is fairly low in ranking. This may be because these measurements are highly correlated with other factors in the state space, such as arterial O_2 pressure (Collins et al. [2015]), that account for its influence on weaning more directly. The limited importance assigned to heart rate and respiratory rate, which can serve as indicators of blood pressure and blood gases, are also likely to be explained by this dependence between vitals.

In terms of demographics, weight and age play a significant role in the weaning policy learnt: weight is likely to influence our sedation policy specifically, as dosages are typically adjusted for patient weight, while age is strongly correlated with a patient's speed of recovery, and hence the time necessary on ventilator support.

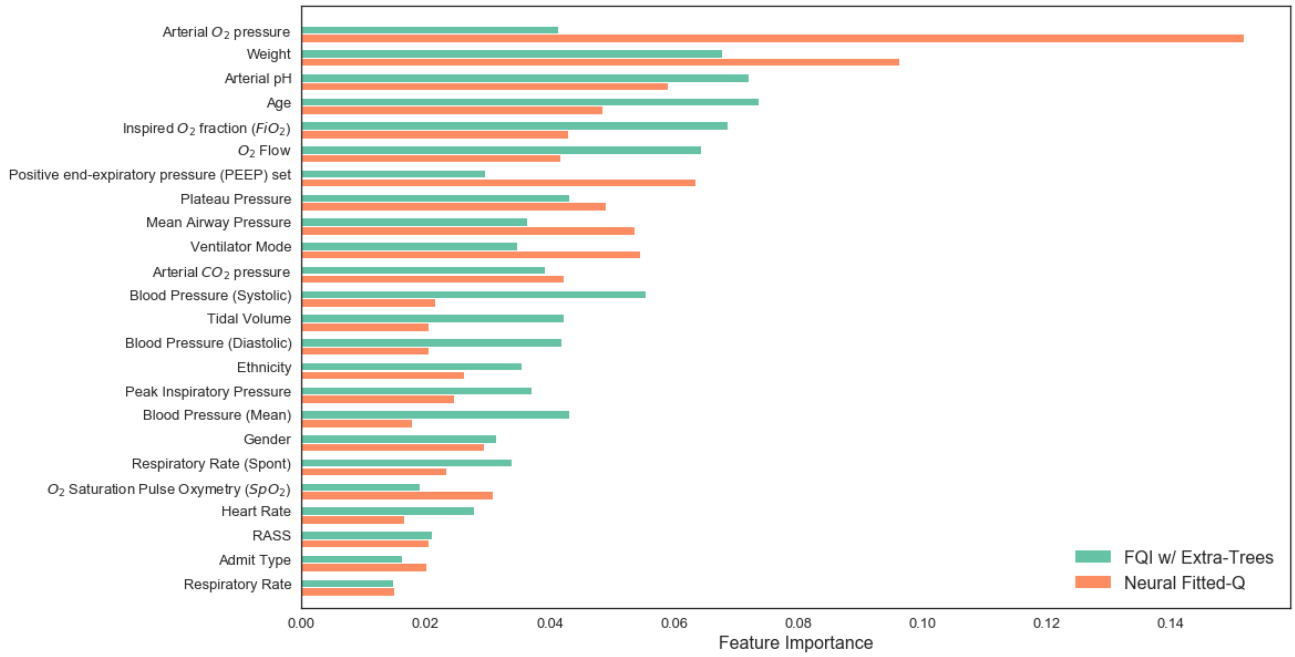


Figure 6: Feature importances (from Gini importance scoring) for policies trained on optimal actions from FQIT & NFQ. The relatively high weighting of indicators *Arterial pH,* FiO_2 and *PEEP set* found is in agreement with typical protocol.

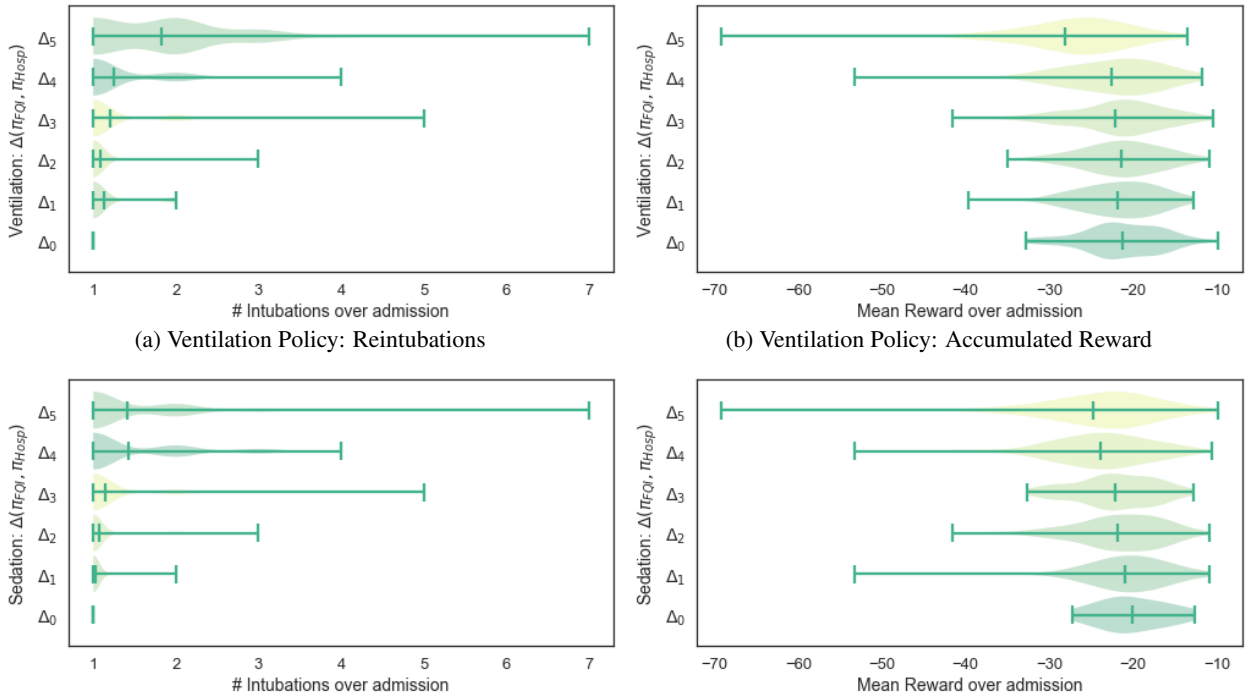
In order to evaluate the performance of the policies learnt, we compare the algorithm's recommendations against the true policy implemented by the hospital. Considering ventilation and sedation separately, the policies learnt with FQIT and NFQ achieve similar accuracies in recommending ventilation (both matching the true policy in roughly 85% of transitions), while FQIT far outperforms NFQ in the case of sedation policy (achieving 58% accuracy compared with just 28%, barely above random assignment of one of four dosage levels), perhaps due to overfitting of the neural network on this task. More data may be necessary to develop a meaningful sedation policy with NFQ.

We therefore concentrate further analysis of policy recommendations to those produced by FQIT. We divide the 664 test admissions into six groups according to the fraction of FQI policy actions that differ from the hospital's policy: Δ_0 comprises admissions in which the true and recommended policies agree perfectly, while those in Δ_5 show the greatest deviation. Plotting the distribution of the number of reintubations and the mean accumulated reward over patient admissions respectively, for all patients in each set (Figures 7a and 7b), we can see that those admissions in set Δ_0 undergo no reintubation, and in general the average number of reintubations increases with deviation from the FQIT policy, with up to seven distinct intubations observed in admissions in Δ_5. This effect is emphasised by the trend in mean rewards across the six admission groups, which serve primarily as an indicator of the regulation of vitals within desired ranges and whether certain criteria were met at extubation: mean reward over a set is highest (and the range lowest) for admissions in which the policies match exactly; mean reward decreases with increasing divergence of the two policies. A less distinct but comparable pattern is seen when grouping admissions instead by similarity of the sedation policy to the true dosage levels administered by the hospital (Figures 7c and 7d).

5 Conclusion

In this work, we propose a data-driven approach to the optimization of weaning from mechanical ventilation of patients in the ICU. We model patient admissions as Markov decision processes, developing novel representations of the problem state, action space, and reward function in this framework. Reinforcement learning with fitted Q-iteration using different regressors is then used to learn a simple ventilator weaning policy from examples in historical ICU data. We demonstrate that the algorithm is capable of extracting meaningful indicators for patient readiness and shows promise in recommending extubation time and sedation levels, on average outperforming clinical practice in terms of regulation of vitals and reintubations.

There are a number of challenges that must be overcome before these methods can be meaningfully implemented in a clinical setting, however: first, in order to generate robust treatment recommendations, it is important to ensure policy invariance to reward shaping: the current methods display considerable sensitivity to the relative weighting of various components of the feedback received after each transition. A more principled approach to the design of the



(a) Ventilation Policy: Reintubations

(b) Ventilation Policy: Accumulated Reward

(c) Sedation Policy: Reintubations

(d) Sedation Policy: Accumulated Reward

Figure 7: Evaluating policy in terms of reward and number of reintubations suggests admissions where actions match our policy more closely are generally associated with better patient outcomes, both in terms of number of reintubations and accumulated reward, which reflects in part the regulation of vitals.

reward function, for example by applying techniques in inverse reinforcement learning (Ng et al. [2000]), can help tackle this sensitivity. In addition, addressing the question of censoring in sub-optimal historical data and explicitly correcting for the bias that arises from the timing of interventions is crucial to fair evaluation of learnt policies, particularly where they deviate from the actions taken by the clinician. Finally, effective communication of the best action, expected reward, and the associated uncertainty, calls for a probabilistic approach to estimation of the Q-function, which can perhaps be addressed by pairing regressors such as Gaussian processes with Fitted Q-iteration.

Possible directions for future work also include increasing the sophistication of the state space, for example by handling long term effects more explicitly using second-order statistics of vitals, applying techniques in inverse reinforcement learning to feature engineering (as in Levine et al. [2010]), or modeling the system as a partially observable MDP, in which observations map to some underlying state space. Extending the action space to include continuous dosages of specific drug types and settings such as ventilator modes and FiO_2 will also facilitate directly actionable policy recommendations. With further efforts to tackle these challenges, the reinforcement learning methods explored here will play a crucial role in helping to inform patient-specific decisions in critical care.

References

Nicolino Ambrosino and Luciano Gabbrielli. The difficult-to-wean patient. *Expert Review of Respiratory Medicine*, 4(5):685–692, 2010.

Hannah Wunsch, Jason Wagner, Maximilian Herlim, David Chong, Andrew Kramer, and Scott D Halpern. Icu occupancy and mechanical ventilator use in the united states. *Critical care medicine*, 41(12), 2013.

Shruti B Patel and John P Kress. Sedation and analgesia in the mechanically ventilated patient. *American journal of respiratory and critical care medicine*, 185(5):486–497, 2012.

Christopher G Hughes, Stuart McGrane, and Pratik P Pandharipande. Sedation in the intensive care setting. *Clin Pharmacol*, 4:53–63, 2012.

James S Krinsley, Praveen K Reddy, and Abid Iqbal. What is the optimal rate of failed extubation? *Critical Care*, 16(1):111, 2012.

Giorgio Conti, Jean Mantz, Dan Longrois, and Peter Tonner. Sedation and weaning from mechanical ventilation: Time for 'best practice' to catch up with new realities? *Multidisciplinary respiratory medicine*, 9(1):45, 2014.

J Goldstone. The pulmonary physician in critical care: Difficult weaning. *Thorax*, 57(11):986–991, 2002. ISSN 0040-6376.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: A reinforcement learning approach. In *2006 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

Yufan Zhao, Donglin Zeng, Mark A. Socinski, and Michael R. Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011. ISSN 1541-0420.

Pablo Escandell-Montero, Milena Chermisi, Jos M. Martnez-Martnez, Juan Gmez-Sanchis, Carlo Barbieri, Emilio Soria-Olivas, Flavio Mari, Joan Vila-Francs, Andrea Stopper, Emanuele Gatti, and Jos D. Martn-Guerrero. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 62(1):47 – 60, 2014. ISSN 0933-3657.

Brett L Moore, Eric D Sinzinger, Todd M Quasny, and Larry D Pyeatt. Intelligent control of closed-loop sedation in simulated icu patients. In *FLAIRS Conference*, pages 109–114, 2004.

R. Padmanabhan, N. Meskin, and W. M. Haddad. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 1–8, Dec 2014.

S. Nemati, M. M. Ghassemi, and G. D. Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2978–2981, Aug 2016.

Martina Mueller, Jonas S Almeida, Romesh Stanislaus, and Carol L Wagner. Can machine learning methods predict extubation outcome in premature infants as well as clinicians? *Journal of neonatal biology*, 2, 2013.

Hung-Ju Kuo, Hung-Wen Chiu, Chun-Nin Lee, Tzu-Tao Chen, Chih-Cheng Chang, and Mauo-Ying Bien. Improvement in the prediction of ventilator weaning outcomes by an artificial neural network in a medical icu. *Respiratory care*, 60(11):1560–1569, 2015.

Yuanyuan Gao, Anqi Xu, Paul Jen-Hwa Hu, and Tsang-Hsiang Cheng. Incorporating association rule networks in feature category-weighted naive bayes model to support weaning decision making. *Decision Support Systems*, 2017.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

Oliver Stegle, Sebastian V. Fallert, David J. C. MacKay, and Søren Brage. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008.

Robert Dürichen, Marco A. F. Pimentel, Lei Clifton, Achim Schweikard, and David A. Clifton. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, 2015.

Marzyeh Ghassemi, Marco A. F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 446–453, 2015.

Li-Fang Cheng, Gregory Darnell, Corey Chivers, Michael Draugelis, Kai Li, and Barbara Engelhardt. Sparse Multi-Output Gaussian Processes for Medical Time Series Prediction. *ArXiv e-prints*, March 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Dirk Ormoneit and Śaunak Sen. Kernel-based reinforcement learning. *Machine learning*, 49(2-3):161–178, 2002.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pages 317–328. Springer, 2005.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Guy W Soo Hoo. Blood gases, weaning, and extubation. *Respiratory Care*, 48(11):1019–1021, 2012. ISSN 0020-1324.

Julie-Ann Collins, Aram Rudenski, John Gibson, Luke Howard, and Ronan ODriscoll. Relating oxygen partial pressure, saturation and content: the haemoglobin–oxygen dissociation curve. *Breathe*, 11(3):194, 2015.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. 2000.

Sergey Levine, Zoran Popovic, and Vladlen Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2010.