FLORIS DEN HENGST
BETA > CS > AI > LEARNING & REASONING
HYBRID INTELLIGENCE CENTRE
F.DEN.HENGST@VU.NL

Guideline-informed reinforcement learning for mechanical ventilation in critical care

AI & Health Meetup 12-10-2023

Mechanical ventilation in intensive care

40% of all ICU patients

~50 000 patients daily in USA

How to ventilate optimally?
How to ventilate safely?

Promises and challenges of reinforcement learning for healthcare

Treatment consists of decisions *over time*

Success of decisions not *immediately* clear

Find *best* treatment rather than *most likely* treatment

Abundance of *data*

Single metric for success

Learn *from scratch*

Mechanical ventilation as a sequential decision-making problem

Based on existing work by Peine et al.

4h time windows

Reward: 90day mortality

States space: 46 features

- demographics
- vital signs
- lab measurements

clustered into 650 states *k*-means clustering

Action spaces: 3 dimensions

- PEEP
- FiO2
- tidal volume

binned into 7^3 = 343 discrete actions

Peine et al., NPJ digital medicine 4 (2021) 32

The ARDSnet protective lung ventilation guideline

Developed in the ARDSnet ARMA trial, extended to all MV patients

Describes allowable treatment actions and target values

$$\varphi_1 \coloneqq FiO_2 \in [0.3, 0.5) \wedge PEEP = 5$$
$$\varphi_2 \coloneqq FiO_2 \in [0.4, 0.6) \wedge PEEP \in [4, 8]$$
$$\varphi_3 \coloneqq FiO_2 \in [0.5, 0.7) \wedge PEEP \in [8,10]$$

…

TABLE 1. SUMMARY OF VENTILATOR PROCEDURES.*

VARIABLE	GROUP RECEIVING TRADITIONAL TIDAL VOLUMES	GROUP RECEIVING LOWER TIDAL VOLUMES
Ventilator mode	Volume assist–control	Volume assist–control
Initial tidal volume (ml/kg of predicted body weight)†	12	6
Plateau pressure (cm of water)	≤50	≤30
Ventilator rate setting needed to achieve a pH goal of 7.3 to 7.45 (breaths/min)	6–35	6–35
Ratio of the duration of inspiration to the duration of expiration	1:1–1:3	1:1–1:3
Oxygenation goal	PaO_2, 55–80 mm Hg, or SpO_2, 88–95%	PaO_2, 55–80 mm Hg, or SpO_2, 88–95%
Allowable combinations of FiO_2 and PEEP (cm of water)‡	0.3 and 5 0.4 and 5 0.4 and 8 0.5 and 8 0.5 and 10 0.6 and 10 0.7 and 10 0.7 and 12 0.7 and 14 0.8 and 14 0.9 and 14 0.9 and 16 0.9 and 18 1.0 and 18 1.0 and 20 1.0 and 22 1.0 and 24	0.3 and 5 0.4 and 5 0.4 and 8 0.5 and 8 0.5 and 10 0.6 and 10 0.7 and 10 0.7 and 12 0.7 and 14 0.8 and 14 0.9 and 14 0.9 and 16 0.9 and 18 1.0 and 18 1.0 and 20 1.0 and 22 1.0 and 24
Weaning	By pressure support; required by protocol when FiO_2≤0.4	By pressure support; required by protocol when FiO_2≤0.4

*PaO_2 denotes partial pressure of arterial oxygen, SpO_2 oxyhemoglobin saturation measured by pulse oximetry, FiO_2 fraction of inspired oxygen, and PEEP positive end-expiratory pressure.

†Subsequent adjustments in tidal volume were made to maintain a plateau pressure of ≤50 cm of water in the group receiving traditional tidal volumes and ≤30 cm of water in the group receiving lower tidal volumes.

‡Further increases in PEEP, to 34 cm of water, were allowed but were not required.

Cruz, et al., Cochrane Database of SystematARDS Network, New England Journal of Medicine 342 (2000) 1301–1308
ic Reviews (2021).
Goligher, Ferguson and Brochard, The Lancet 387 (2016) 1856–1866.
Fernando et al., Chest 159 (2021) 606–618.

Safe reinforcement learning – policy level

We use an action filter:

$$\mathcal{A}_s \subseteq \mathcal{A}: \{a \in \mathcal{A} \mid \vee_i\, a \vDash \varphi_i\}$$

we can apply the safety filter *after training*

$$\pi_{\mathcal{C}}(a|s) = \begin{cases} \dfrac{\pi(a|s)}{\sum_{a' \in \mathcal{A}_s} \pi(a'|s)} & \text{if } a \in \mathcal{A}_s \\ 0 & \text{otherwise} \end{cases}$$

Safe reinforcement learning – Q-function

We use an action filter:

$$\mathcal{A}_s \subseteq \mathcal{A}: \{a \in \mathcal{A} \mid \vee_i \; a \vDash \varphi_i\}$$

we can apply the safety filter *during training*

$$\hat{Q}(s,a) \leftarrow \begin{cases} \hat{Q}(s,a) + \alpha\left[r + \gamma \max_{a' \in \mathcal{A}_s} \hat{Q}(s',a') - \hat{Q}(s,a)\right] & \text{if } a \in \mathcal{A}_s \\ -\infty & \text{otherwise} \end{cases}$$

Guideline-informed reward shaping

$$\varphi_1 \coloneqq Pplat \leq 30$$
$$\varphi_2 \coloneqq pH \in (7.2, 7.5)$$
$$\ldots$$

Desirability function $F: \mathcal{X} \rightarrow \{0, 1\}$ for states

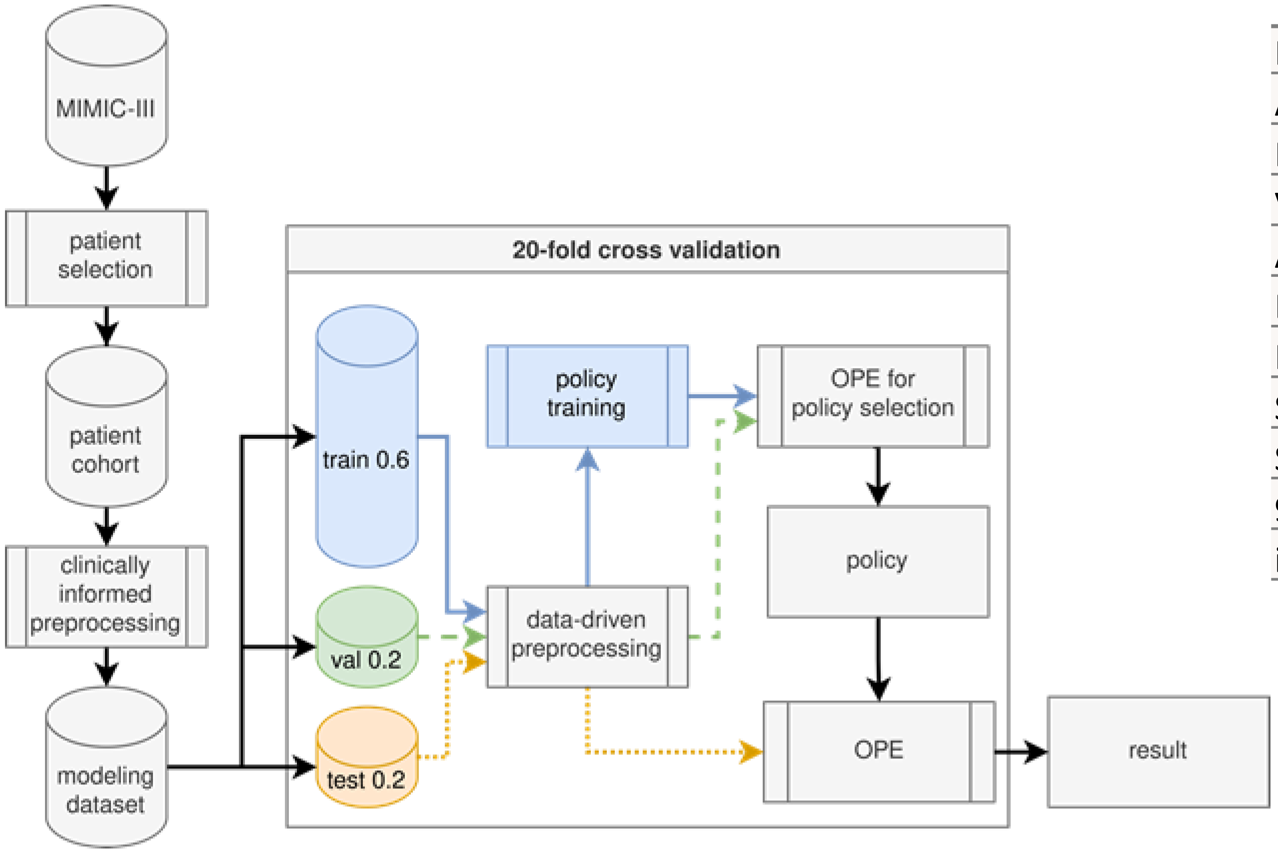
$$F \subseteq \mathcal{X}: \{x \in \mathcal{X} \mid \wedge_i \, x \vDash \varphi_i\}$$

Reward shaping

$$R'(x, a, x') = R(x, a, x') + \gamma c F(x') - c F(x)$$

with some scalar $c \in \mathbb{R}^+$ to balance original and shaping rewards

Retrospective study design



Number of ICUs	5
Acquisition timespan	2001-2012
Patients	7659
Ventilation events	8799
Age	65.67 (53.19-76.44) years
Body weight	86.24 ± 24.89 kg
Ideal body weight	63.38 ± 12.93 kg
Sex, female	3813 (43.33%)
Sex, male	4986 (56.67%)
90-day mortality	34.50%
in-hospital mortality	25.73%

Retrospective evaluation with off-policy evaluation

Inverse propensity scoring

PH-WIS

adjust for differences between the learned and clinician's solution

depends only on data

high variance

Model-based

FQE

a model is learned to estimate performance of the learned solution

requires proper estimation

low variance

Hybrid

combination of the previous

Doroudi, Thomas and Brunskill, IJCAI, 2018

Evaluating the approaches

	unconstrained	safe, after learning	safe, during learning
reinforcement learning / stochastic	●	●	●
reinforcement learning / deterministic	●	●	●
observed	●		
imitation learning	●	●	

expected rewards

- model-based
- inverse propensity
- hybrid

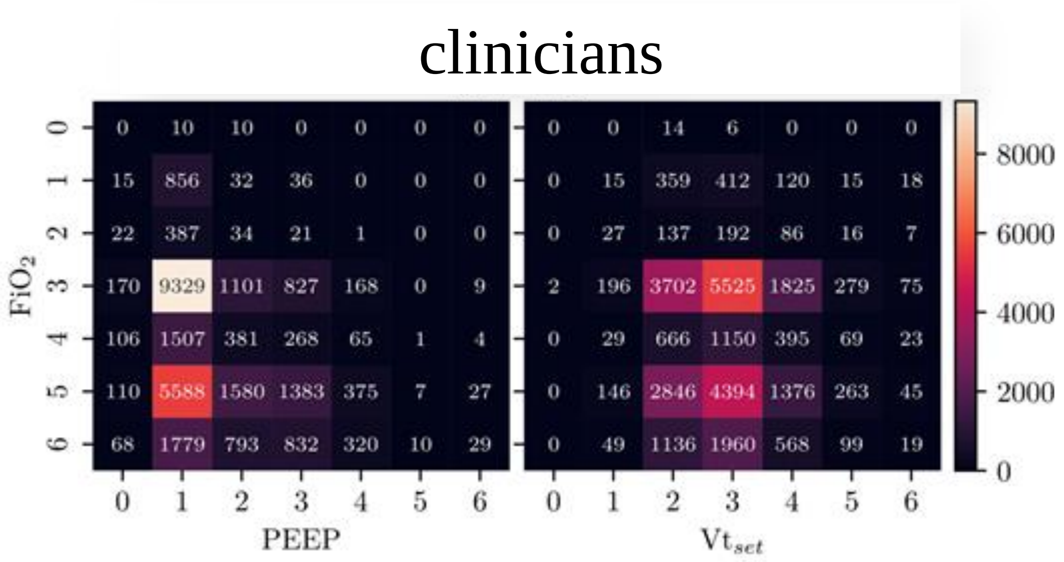
safety

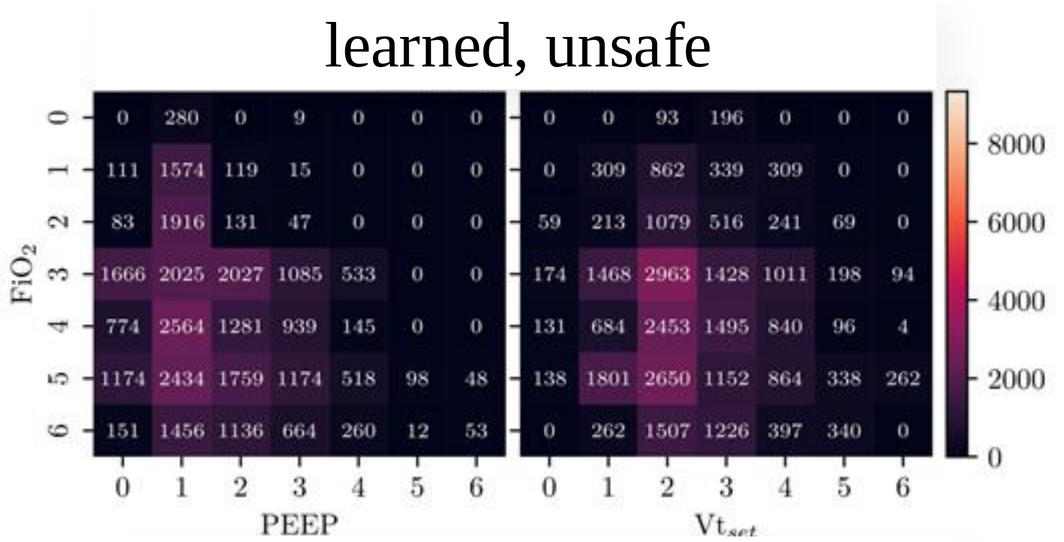
- probability of unsafe action

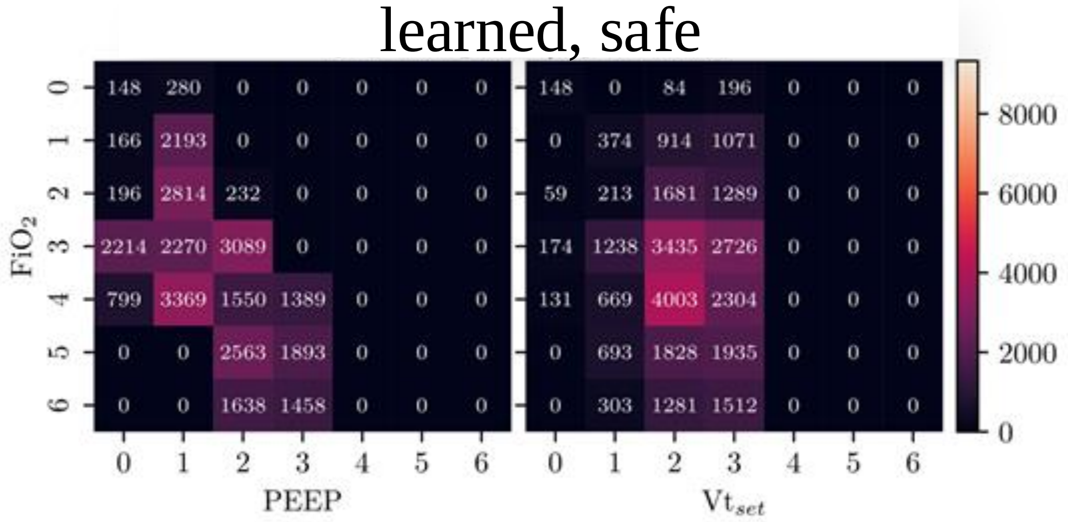
design effect

- effective sample size

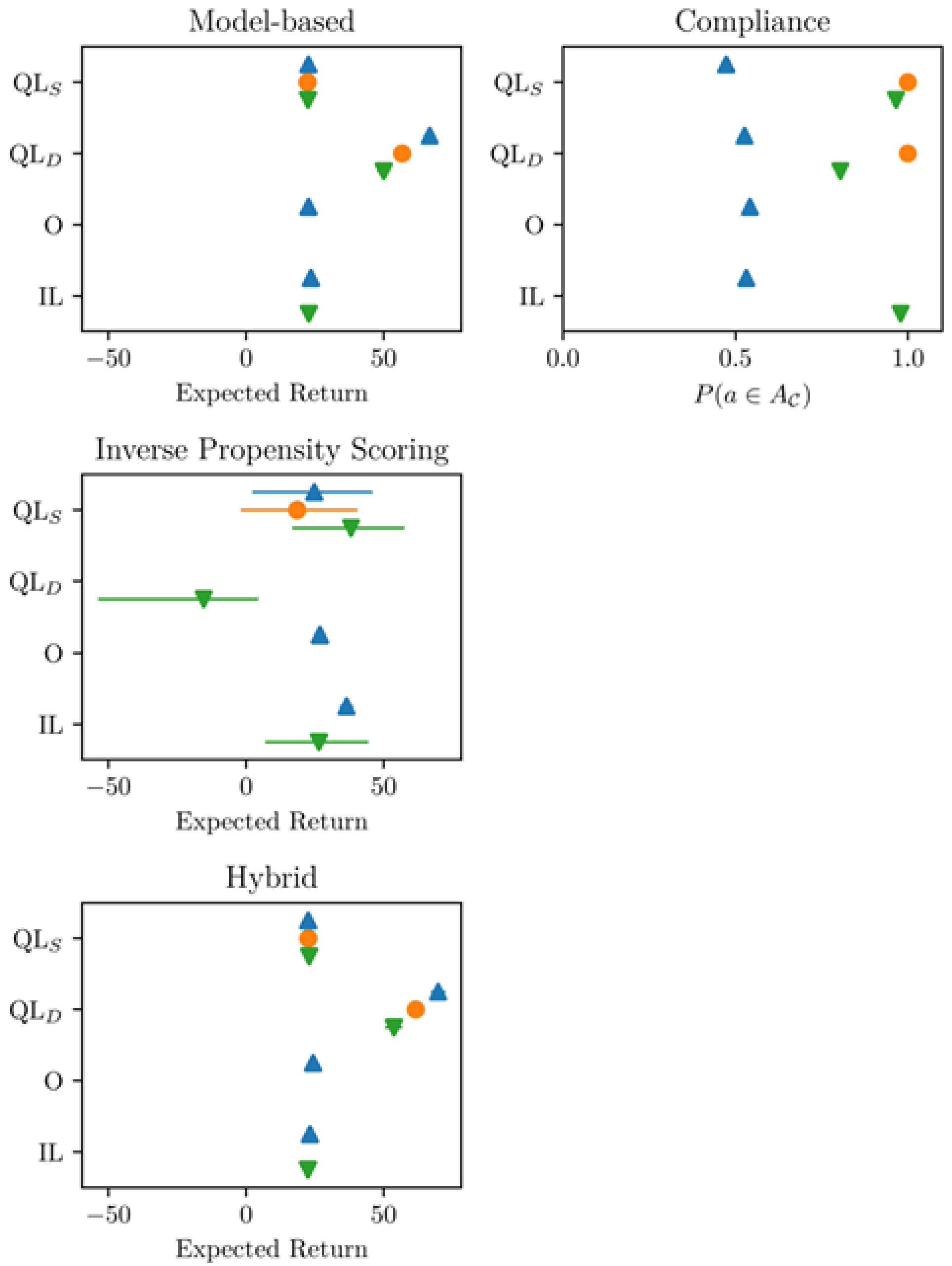
Learned policies are more varied







Comparable, but safer



Discussion

A reinforcement learning framework to optimize clinical decision-making with observational data and existing knowledge

strict guideline compliance

outperforms clinicians in a model-based evaluation

problematic model-free evaluation: differences in decision-making

user-specified compliance rate

different use cases

multiple objectives

Thank you

Martijn
Otten

Paul
Elbers

Vincent
François-Lavet

Mark
Hoogendoorn

Frank
van Harmelen

VRIJE UNIVERSITEIT AMSTERDAM

Amsterdam UMC
Universitair Medische Centra

Questions

Appendix

Inverse Propensity Scoring

returns are weighted to adjust for differences evaluation and behavior policies

Model-based

a model is learned to estimate Q values of the behavior policy

Hybrid

Combination of previous

per-horizon
weighted importance sampling

$$\rho_t^{tr} = \prod_{i=1}^{t} \frac{\pi_e(a_i^{tr}|s_i^{tr})}{\pi_b(a_i^{tr}|s_i^{tr})}.$$

$$W_l = \frac{|\{tr_i|T_i = l\} \in D|}{n}$$

$$\hat{V}_{\pi_e}^{\mathrm{PHWIS}}(D) = \sum_{l \in \mathcal{L}} W_l \sum_{\{tr_i|T_i=l\}} \sum_{t=0}^{T_i-1} \frac{\rho_t^{(i)}}{\sum_{\{tr_i|T_i=l\}} \rho_t^{(i)}} \gamma^t \, r_t^{(i)}$$

fitted Q evaluation

$$\widehat{Q}_k = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{\tilde{T}} (\widehat{Q}_{k-1}(x_t^i, a_t^i; \theta) - y_t^i)^2$$

$$y_t^i \equiv r_t^i + \gamma \mathbb{E}_{\pi_e} \widehat{Q}_{k-1}(x_{t+1}^i, \cdot; \theta)$$

per-horizon
weighted doubly robust

C. Voloshin, H. M. Le, N. Jiang, Y. Yue, Empirical study of off-policy policy evaluation for reinforcement learning. Thirty-fifth Conference on Neural Information Processing Systems, 2021

The intensive care unit

closely monitored

public datasets available

highly controlled

life or death scenarios

intensive care

Space	Variable	Guideline	Constraint	#
State	Pplat	≤ 30	≤ 30	φ_1
	pH	$7.3 - 7.45$	$\in (7.2, 7.5)$	φ_2
	RR	$6 - 35$	≤ 35	φ_3
	SpO_2	$88 - 95\%$	≥ 88	φ_4
Action	Vt_{set}	6 (initial)	≤ 8.5	φ_5
	FiO_2	0.3 and 5	$FiO_2 \in [0.3, 0.5) \quad \wedge \quad$ PEEP$= 5$	φ_6
	and	0.4 and 5		
	PEEP	0.4 and 8	$FiO_2 \in [0.4, 0.6) \quad \wedge \quad$ PEEP$\in [4, 8]$	φ_7
		0.5 and 8		
		0.5 and 10	$FiO_2 \in [0.5, 0.7) \quad \wedge \quad$ PEEP$\in [8, 10]$	φ_8
		0.6 and 10		
		0.7 and 10	$FiO_2 \in [0.7, 0.8) \quad \wedge$ PEEP$\in [10, 14]$	φ_9
		0.7 and 12		
		0.7 and 14		
		0.8 and 14	$FiO_2 \in [0.8, 0.9) \quad \wedge \quad$ PEEP$= 14$	
		0.9 and 14	$FiO_2 \in [0.9, 1.0) \quad \wedge$ PEEP$\in [14, 18]$	φ_{10}
		0.9 and 16		
		0.9 and 18		
		1.0 and 18	$FiO_2 = 1.0 \quad \wedge$ PEEP$\in [18, 24]$	φ_{11}
		1.0 and 20		
		1.0 and 22		
		1.0 and 24		

	Variable	Imputation window (h)	% Missing Initial	1^{st} Step	2^{nd} Step
demographic	Age	–	0.0	0.0	0.0
	IBW		16.8	16.8	
	Height		16.8	16.8	
	Weight		14.4	14.4	
	ICU readmission		0.0	0.0	
	Elixhauser-vanWalraven		77.2	77.2	
vital signs	SOFA	24	0.0	0.0	0.0
	SIRS	24	0.0	0.0	
	GCS	$*$	19.0	1.5	
	HR	$*$	1.4	0.6	
	SysBP	$*$	2.5	0.6	
	MeanBP	$*$	1.9	0.6	
	DiasBP	$*$	2.5	0.6	
	ShockIndex	$*$	3.2	0.7	
	RR	$*$	1.8	0.6	
	SpO_2	$*$	2.2	0.6	
	TempC	$*$	9.4	1.2	
action	PEEP	8	33.7	25.3	0.0
	FiO_2		26.1	16.6	
	Vt_{set}		33.8	25.4	
other	IV	8	14.1	5.7	0.0
	Urine output	8	14.5	11.6	
	Fluid Balance	8	3.6	2.4	
	Plateau Pressure	8	80.1	7.8	
	Vasopressors (dosage)	24	88.4	74.8	
	PaO_2/FiO_2 ratio	$*$	98.3	55.7	

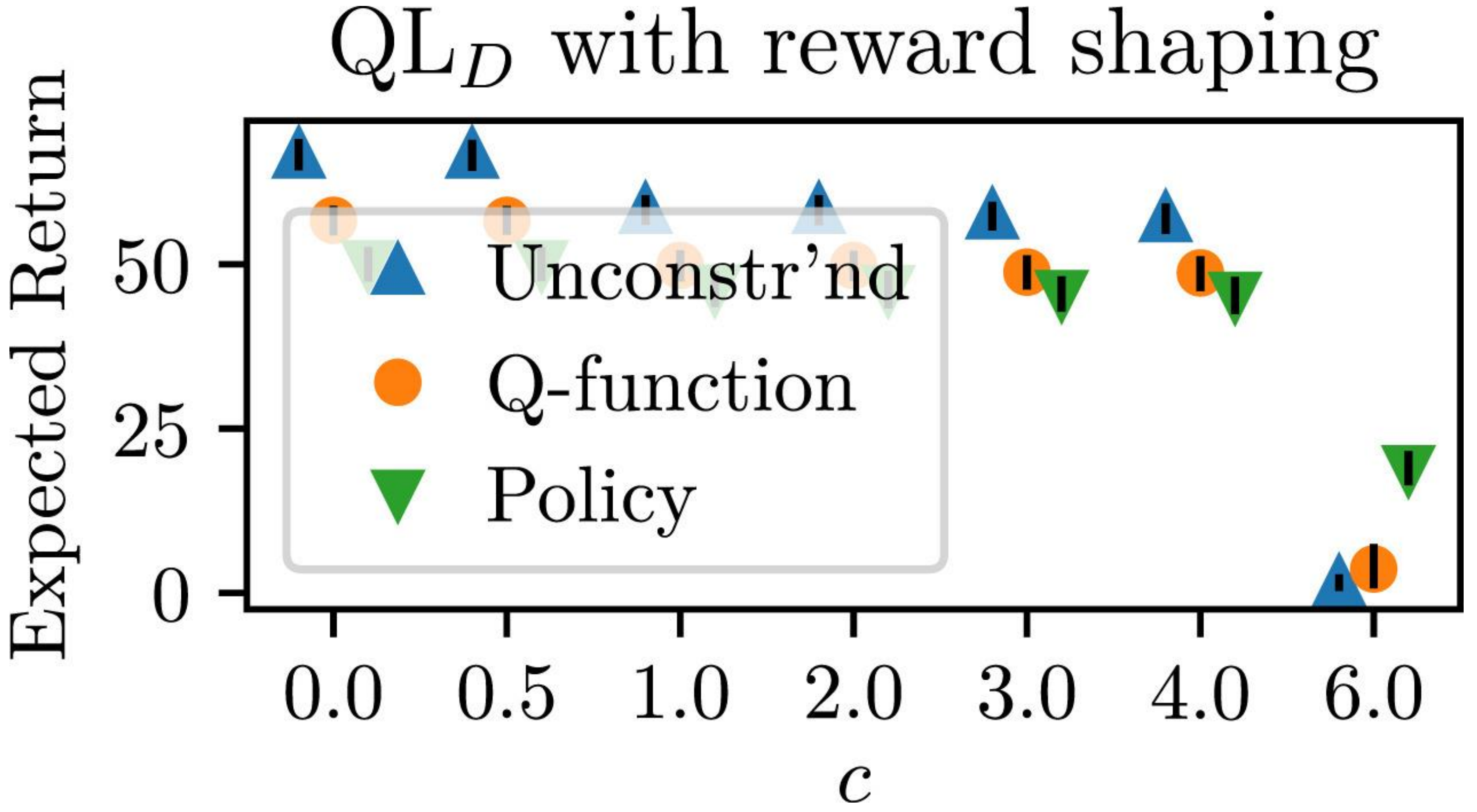
Variable		Imputation window (h)	% Missing		
			Initial	1^{st} Step	2^{nd} Step
lab measurements	Potassium	$*$	95.4	4.2	0.0
	Sodium	$*$	95.5	3.9	
	Chloride	$*$	95.5	3.4	
	Glucose	$*$	95.7	4.8	
	BUN	$*$	95.5	2.6	
	Creatinine	$*$	95.5	2.6	
	Magnesium	$*$	95.5	7.0	
	Calcium	$*$	95.9	11.3	
	Ionized Calcium	8	96.4	56.8	
	Calculated Carbon Dioxide†	$*$	83.2	9.7	
	Bilirubin	$*$	94.0	42.4	
	Albumin	$*$	98.4	51.4	
	Hemoglobin	$*$	98.9	3.0	
	WBC	$*$	95.8	2.9	
	Platelet	$*$	95.6	2.5	
	PTT	$*$	96.3	8.5	
	PT	$*$	96.2	8.0	
	INR	$*$	96.2	8.0	
	PH	$*$	93.9	8.9	
	PaO_2	$*$	97.8	45.8	
	$PaCO_2$†	$*$	94.4	12.7	
	Base Excess	$*$	94.5	13.0	
	Bicarbonate	$*$	95.6	3.4	
	Lactate	$*$	96.3	21.2	

Variable	#	Range	Variable	#	Range	Variable	#	Range
	1	$[0, 2.5)$		1	$[0, 5)$		1	$[20, 30)$
	2	$[2.5, 5)$		2	$[5, 7)$		2	$[30, 35)$
	3	$[5, 7.5)$		3	$[7, 9)$		3	$[35, 40)$
Vt_{set}	4	$[7.5, 10)$	PEEP	4	$[9, 11)$	FiO_2	4	$[40, 45)$
	5	$[10, 12.5)$		5	$[11, 13)$		5	$[45, 50)$
	6	$[12.5, 15)$		6	$[13, 15)$		6	$[50, 55)$
	7	$[15, \infty)$		7	$[15, \infty)$		7	$[55, \infty)$

Table A.6: Action discretization: all actions variables were binned into seven bins. Each combination of bins for all variables was then mapped to a single action, resulting in a total of $7^3 = 343$ discrete actions.

		PHWIS		PHWDR	
Algorithm	Compliance	Statistic	p-value	Statistic	p-value
IL	Unconstrained	209.0	0.999	16.0	0.000
	Policy	107.0	0.536	16.0	0.000
QL_D	Unconstrained	–	–	210.0	1.0
	Policy	0.0	0.125	210.0	1.0
	Q-function	–	–	210.0	1.0
QL_S	Unconstrained	93.0	0.337	16.0	0.000
	Policy	80.0	0.184	14.0	0.000
	Q-function	128.0	0.806	25.0	0.000

Table A.7: One-tailed significance test results for the listed policy having a *lower* mean expected return than observed in the test set obtained with Wilcoxon's signed-rank test. Results for QL_D Unconstrained and QL_D Q-function are missing due to an expected sample size of zero.


QL$_D$ with reward shaping
Expected Return
0
25
50
0.0
0.5
1.0
2.0
3.0
4.0
6.0
c
Unconstr'nd
Q-function
Policy

4.1. Formalisation of guidelines

We consider a set of l variables $\mathcal{V} : \{\nu_1, \dots, \nu_l\}$ to describe patient states and treatment decisions and a finite set of m ranges to describe allowable ranges $\mathcal{R} = \left\{\nu_1 \in [v_{\min}^{(1)}, v_{\max}^{(1)}], \dots, \nu_1 \in [v_{\min}^{(i)}, v_{\max}^{(i)}], \dots, \nu_j \in [v_{\min}^{(m)}, v_{\max}^{(m)}]\right\}$ for these variables. We consider each clause φ as a subset of the power set of ranges $\varphi \subseteq 2^{\mathcal{R}}$. We require that all values fall within the provided bounds in φ to comply to that clause. Formally, a set of measurements $\{\nu_1 = v_1, \dots, \nu_n = v_n\} \models \varphi \iff \bigwedge_{j=0}^{|\varphi|} \left(v_i \in \left[v_{\min}^{(j)}, v_{\max}^{(j)}\right] \vee \nu_i \neq \nu_j\right)$ where $|\varphi|$ denotes the number of ranges in φ. Note that multiple ranges can be assigned to a single variable ν within the guideline $\mathcal{R}$ but that these ranges should overlap.