ORIGINAL RESEARCH **Open access**

Reinforcement Learning Framework for Dynamic Optimization of Extracorporeal Membrane Oxygenation Settings Using Real-Time Blood Gas, Hemodynamic, and Pump Flow Measurements

Yuki Yamamoto[1*], Kenji Ito[1]

Abstract

Extracorporeal membrane oxygenation (ECMO) is used to support patients with severe cardiac or respiratory failure, requiring constant manual adjustments of pump flow, sweep gas flow, and oxygen fraction. However, current ECMO management lacks a real-time optimization system tailored to individual patient needs. This manuscript proposes an offline reinforcement learning framework for dynamic ECMO optimization, utilizing real-time measurements of blood gases, hemodynamics, and pump flow. The framework includes a state encoder for various patient data, an action space for adjustments to ECMO settings, and a reward function that balances oxygenation, hemodynamic support, and complication avoidance. A safety shield filters unsafe recommendations before clinician review. The system aims to provide personalized, proactive, and safety-constrained ECMO management, with the goal of guiding future research validation rather than claiming experimental results.

Keywords Reinforcement learning, Extracorporeal membrane oxygenation, Offline reinforcement learning, Blood gas monitoring, Hemodynamic monitoring, Pump flow

*Correspondence:
Yuki Yamamoto
yuki.yamamoto@gmail.com
[1] Department of Healthcare AI Systems, Osaka University, Osaka, Japan

Introduction

Extracorporeal membrane oxygenation is used in severe respiratory failure, cardiogenic shock, extracorporeal cardiopulmonary resuscitation, and other states where native cardiopulmonary function is insufficient to sustain life. Venovenous ECMO primarily supports gas exchange, whereas venoarterial ECMO provides circulatory support in addition to oxygen delivery, making the consequences of pump flow, sweep gas flow, and FiO_2 adjustments highly context dependent [1, 2]. Contemporary guidance emphasizes structured management, physiologic targets, and complication surveillance, but bedside titration remains manual and reactive. This creates an opportunity for decision-support systems that can integrate changing physiologic signals into patient-specific recommendations rather than relying only on periodic reassessment [3, 4].

ECMO generates continuous and intermittent data streams that are clinically rich but difficult to synthesize in real time. Blood gases provide information about pH, PaO_2, $PaCO_2$, bicarbonate, lactate, and SaO_2, while hemodynamic monitoring captures arterial pressure, heart rate, central venous pressure, and venous oxygenation status [1, 2]. Pump flow, sweep gas flow, FiO_2, pump speed, and circuit pressures add device-specific information about oxygen delivery, decarboxylation, vascular drainage, oxygenator function, and mechanical stress [4, 5]. Despite this information density, most ECMO management pathways do not implement a systematic optimization framework that

learns dynamic setting-response relationships from historical trajectories [3, 6].

Offline reinforcement learning is attractive for ECMO because it can learn decision policies from historical data without requiring unsafe exploration at the bedside. Prior critical-care RL work has demonstrated policy-learning approaches for ventilation, sepsis therapy, vasopressor timing, sedation, and corticosteroid strategies, showing how sequential clinical decisions can be framed as optimization problems under uncertainty [7-13]. However, ECMO differs from many critical-care interventions because device settings directly affect gas exchange, circulatory support, hemolysis risk, bleeding risk, thrombosis risk, and weaning readiness. Therefore, an ECMO-focused RL framework must be conservative, interpretable, safety constrained, and compatible with human clinical oversight [3, 14, 15].

This manuscript proposes a conceptual framework for offline reinforcement learning to optimize ECMO settings using real-time blood gas, hemodynamic, and pump flow measurements. The framework is not an experimental study and does not report simulated or clinical performance results. Instead, it defines the state representation, action space, transition assumptions, offline RL architecture, and safety principles required for a future ECMO decision-support system [3, 4, 16]. The roadmap proceeds from ECMO physiology and monitoring to MDP formulation, offline policy learning, and a real-time clinician-facing recommendation workflow.

Background

ECMO physiology

The physiologic effect of ECMO depends on the interaction between native cardiopulmonary function, cannulation strategy, circuit performance, and device settings. Sweep gas flow primarily controls carbon dioxide removal by altering gas exchange across the membrane lung, while FiO_2 influences oxygen transfer across the oxygenator and pump flow determines the volume of blood exposed to extracorporeal gas exchange [1, 2]. In venoarterial ECMO, pump flow also contributes to systemic perfusion and may interact with ventricular loading, arterial pressure, and differential hypoxemia, whereas in venovenous ECMO, recirculation, native lung function, and cardiac output shape oxygen delivery. These physiologic relationships make ECMO titration a sequential control problem rather than a single static prescription [3, 5].

Real-time monitoring

Real-time ECMO management requires the integration of arterial blood gas values, hemodynamic variables, ventilator context, device parameters, and circuit measurements. Blood gas variables such as pH, PaO_2, $PaCO_2$, bicarbonate, lactate, and SaO_2 describe the adequacy of oxygenation, ventilation, perfusion, and metabolic recovery, while MAP, CVP, $ScvO_2$, heart rate, and vasoactive support contextualize circulatory stability [1, 2]. Pump flow, pump speed, drainage pressure, return pressure, transmembrane pressure, and oxygenator performance provide device-level signals that can indicate inadequate support, excessive suction, oxygenator dysfunction, or thrombotic burden [4, 5]. A learning system for ECMO optimization should therefore represent the patient and the circuit as a coupled physiologic-technological state rather than as separate monitoring domains [3, 6].

Current management protocols

Current ECMO management is typically protocol informed but clinician directed, with titration based on blood gas targets, hemodynamic goals, ventilator strategy, anticoagulation status, and daily reassessment of readiness to wean. Guidelines for adult venovenous and venoarterial ECMO emphasize structured care, surveillance for bleeding, thrombosis, hemolysis, neurologic injury, and device complications, but they do not prescribe a continuously optimized control policy for setting adjustment [1, 2]. Weaning is commonly approached through goal-directed reductions in support, trials of lower sweep gas or flow, and assessment of whether native cardiopulmonary function can maintain stability. This workflow is clinically grounded but reactive, creating a gap between high-frequency data generation and proactive individualized decision support [3, 4].

Offline reinforcement learning

Offline reinforcement learning learns from fixed datasets of previous decisions and outcomes, making it suitable for critical-care domains where online exploration would be unethical or dangerous. Methods such as conservative Q-learning, implicit Q-learning, and behavior-regularized actor-critic strategies are designed to reduce extrapolation beyond the support of observed clinical practice, a core concern in high-stakes settings [10, 14, 15]. Critical-care studies in ventilation, sepsis, sedation, corticosteroids, and

vasopressor initiation show that RL can represent treatment as a sequence of state-dependent decisions rather than as isolated prescriptions [7-9, 11-13, 17]. For ECMO, this offline orientation is essential because unsafe trial-and-error changes in pump flow, sweep gas flow, or FiO_2 could immediately compromise oxygen delivery, decarboxylation, hemodynamics, or circuit safety [3, 16].

Framework Overview

High-level architecture

The proposed framework converts real-time ECMO monitoring data into a structured patient-circuit state, passes that state through an offline-trained RL policy, and generates a recommended incremental action. The recommendation may include increasing or decreasing pump flow, sweep gas flow, or FiO_2, after which a safety shield checks hard clinical limits, rate-of-change constraints, and predicted risk signals before the recommendation reaches the bedside team [1-3]. The system is not designed for autonomous control; rather, it functions as clinician-facing decision support that explains why a recommendation is aligned with gas exchange, hemodynamic stability, and complication avoidance. This architecture follows the broader critical-care RL principle that algorithmic policies should augment expert judgment rather than replace accountability in safety-critical environments [10, 14, 15].

Figure 1 presents the proposed safety-constrained offline reinforcement learning architecture for transforming real-time ECMO monitoring data into clinician-reviewed setting recommendations.

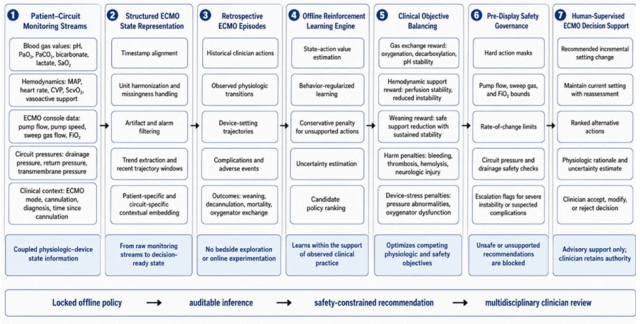

Figure 1. Safety-Constrained Offline Reinforcement Learning Architecture for Dynamic ECMO Setting Optimization

Core assumptions

The framework assumes access to historical ECMO data covering the period from cannulation to decannulation, death, transplant, durable support, or transfer. Each trajectory should include time-stamped device settings, blood gas values, hemodynamic variables, medication context, ventilation variables, circuit measurements, adverse events, and outcomes such as survival, weaning, bleeding, thrombosis, hemolysis, neurologic injury, and oxygenator exchange [6, 18-20]. Training occurs entirely offline, and the deployed model does not improve itself by experimenting on patients. Safety constraints, action bounds, exclusion criteria, and escalation rules must be specified before deployment because ECMO decisions involve immediate physiologic consequences [1-3].

Design principles

The framework is governed by four design principles: safety-first recommendation, data-efficient offline learning, interpretable action justification, and low-latency bedside integration. Safety-first design follows from ECMO guidance and registry-based risk modeling showing that complications such as neurologic injury, thrombosis, bleeding, and early mortality are central determinants of outcomes [1, 2, 18-22]. Data efficiency is required because high-quality ECMO datasets are smaller and more heterogeneous than general ICU datasets, while interpretability is necessary for clinician acceptance of recommendations that modify life-support settings. Low-latency inference is important because blood gas abnormalities, hypotension, drainage insufficiency, and circuit pressure changes may require timely response rather than delayed retrospective review [3, 4].

MDP Formulation

State space

The Markov decision process state should encode both patient physiology and extracorporeal circuit function at each decision time. Continuous variables include MAP, heart rate, CVP, $ScvO_2$, pH, PaO_2, $PaCO_2$, bicarbonate, lactate, SaO_2, pump flow, pump speed, sweep gas flow, FiO_2, drainage pressure, return pressure, transmembrane pressure, ventilator settings, vasoactive support, anticoagulation indicators, time since cannulation, and recent trends [1-3, 5]. The state should also include contextual variables such as ECMO mode, cannulation

configuration, diagnosis, age, organ dysfunction, bleeding status, oxygenator age, and prior adverse events because identical settings may have different implications across patients. This representation allows the policy to recommend setting changes based on trajectories rather than isolated threshold violations [6, 18, 21].

Table 1 clarifies how ECMO physiology can be translated into reinforcement learning states, actions, rewards, and safety constraints without reducing bedside management to isolated numeric thresholds.

Table 1. Translating ECMO Physiology into Reinforcement Learning Design Elements

ECMO management domain	Physiologic meaning	RL design representation	Req int
Oxygenation control	Adequacy of arterial oxygen delivery through native lung and extracorporeal support	PaO_2, SaO_2, FiO_2, pump flow, ECMO mode, ventilator context	Avo FiO or u flow
Decarboxylation control	Carbon dioxide removal through sweep gas and membrane lung exchange	$PaCO_2$, pH, bicarbonate, sweep gas flow, recent blood gas trends	Pre swe dest ba
Circulatory support	Perfusion adequacy, especially in VA ECMO	MAP, CVP, $ScvO_2$, heart rate, vasoactive support, lactate, pump flow	recor th oxyg w pe v
Circuit function	Mechanical and oxygenator performance of the extracorporeal circuit	Pump speed, drainage pressure, return pressure, transmembrane	Blo ass suc pres or dy
		pressure, oxygenator age	
Complication risk	Bleeding, thrombosis, hemolysis, neurologic injury, oxygenator failure	Adverse-event indicators, pressure trends, anticoagulation context, prior complications	Pena phys tha maj
Weaning readiness	Recovery of native cardiopulmonary function and ability to tolerate lower support	Stable gas exchange, improving hemodynamics, decreasing support needs, time trends	Re susta af r

Action space

The action space consists of clinically plausible changes in ECMO settings rather than unconstrained absolute prescriptions. Pump flow actions may represent incremental changes such as reductions or increases within a narrow range, sweep gas actions may adjust decarboxylation support, and FiO_2 actions may alter oxygenator oxygen fraction while respecting minimum and maximum bounds [1-3]. The action space can be discrete, with bins for no change and small upward or downward adjustments, or continuous, with bounded deltas for pump flow, sweep gas flow, and FiO_2. A discrete action space may improve interpretability and off-policy evaluation reliability, whereas a continuous action space may better represent fine titration if sufficient historical data support exists [7, 8, 23].

Transition dynamics

The transition function represents how patient-circuit states evolve after an ECMO setting change, but this function is unknown and must be learned from historical observations. Model-free RL can estimate action values directly from trajectories, while model-based RL can learn physiologic transition approximations and use them for counterfactual simulation or policy evaluation [10, 14, 23]. In ECMO, transition dynamics are complicated by delayed blood gas response, changes in native lung or cardiac function, anticoagulation, transfusion, ventilation strategy, vasoactive therapy, and circuit aging. A conservative framework should therefore avoid aggressive extrapolation and treat rarely

observed combinations of state and action as uncertain or unsafe unless supported by sufficient historical evidence [3, 4, 16].

Offline RL Architecture

Algorithm selection

The preferred algorithmic family is conservative offline reinforcement learning because ECMO policy learning must avoid recommending actions that are poorly represented in the historical dataset. Conservative Q-learning, implicit Q-learning, and behavior-regularized actor-critic methods are appropriate candidates because they penalize out-of-distribution actions and reduce extrapolation error, which is a major threat in retrospective clinical RL [10, 14, 15]. Prior critical-care RL studies illustrate the importance of conservative evaluation and reward shaping in ventilation, sepsis, vasopressor, and sedation domains, where learned policies may otherwise appear optimal because of data artifacts rather than clinical safety [7-9, 11, 12, 17]. For ECMO, the algorithm should prefer actions near the clinician behavior distribution unless strong evidence supports a safer or more physiologically beneficial alternative [3, 16].

Network architecture

The network architecture may use a deep Q-network for discrete actions or an actor-critic structure for bounded continuous action recommendations. Inputs would include approximately twenty to thirty state variables representing blood gas values, hemodynamics, ventilator context, pump parameters, circuit pressures, trends, and time on support, with two or three hidden layers sufficient for a conceptual baseline architecture before considering recurrent or transformer models [3, 7, 8]. Temporal modeling is important because ECMO physiology depends on trends, delayed response, and cumulative exposure, so recurrent encoders or sequence windows may be useful if data density supports them. The output should be clinically interpretable, such as Q-values for candidate actions or ranked recommendations with uncertainty estimates, rather than an opaque command to change life-support settings [4, 10, 14].

Training procedure

Training should use historical ECMO episodes with known trajectories and outcomes, including survival, successful decannulation, duration of support, oxygenator exchange, hemolysis, thrombosis, bleeding, neurologic injury, and treatment escalation. The dataset should be split by patient rather than by time point to prevent leakage, and preprocessing should harmonize irregular blood gas measurements, continuous monitor data, medication changes, and device logs into aligned decision intervals [3, 6, 18-22]. Pessimistic regularization should be applied so that the learned policy assigns lower value to unsupported actions and uncertain transitions. Before clinical use, the policy must remain locked, audited, and evaluated offline against historical clinician decisions, rule-based protocols, and safety-violation criteria rather than being adapted through uncontrolled online learning [10, 14, 15].

Reward Design

Primary objectives

The primary reward should encode physiologic improvement rather than merely reproduce historical clinician behavior. Positive reward is assigned when ECMO settings move the patient toward target oxygenation, target decarboxylation, and acid-base stability, such as PaO_2 within an acceptable oxygenation range, $PaCO_2$ near physiologic values, and pH normalization without excessive device support [1-3]. Related RL work in ventilation and anesthesia demonstrates that reward functions must reflect clinically meaningful physiologic endpoints rather than isolated device adjustments [7-9, 24]. In ECMO, the reward should therefore balance immediate gas exchange correction with avoidance of unnecessary pump flow, sweep gas flow, or FiO_2 escalation.

Penalties

Negative reward should penalize trajectories associated with hemolysis, bleeding, thrombosis, oxygenator dysfunction, circuit pressure abnormalities, arrhythmia, neurologic injury, and hemodynamic deterioration. These penalties are essential because a policy that optimizes blood gas values alone could recommend excessive flow, abrupt sweep changes, or prolonged high oxygen exposure while increasing complication risk [1, 2, 18-20]. Critical-care RL studies in ventilation, sedation, sepsis, and vasopressor therapy show that reward design must explicitly account for harm, delayed outcomes, and competing clinical objectives [11, 12, 25-27]. In this framework, complication penalties should be weighted conservatively so that marginal physiologic gains do not override major safety risks.

Weaning reward

A weaning reward should encourage safe reductions in ECMO support when the patient demonstrates stable gas exchange, improving hemodynamics, and adequate native cardiopulmonary recovery. The reward should not simply favor lower settings, because premature reductions in pump flow, sweep gas flow, or FiO_2 may cause rebound hypoxemia, hypercapnia, acidosis, circulatory collapse, or emergency recannulation-like rescue escalation [1, 2]. Prior sepsis and critical-care RL studies show that policies must recognize longitudinal recovery patterns rather than reward isolated short-term changes [13, 28-30]. Therefore, successful weaning reward should be granted only when reduction in support is followed by sustained stability and eventual decannulation or transition to a lower-risk support state.

Safety Constraints

Hard actin masks

Hard action masks define non-negotiable limits that prevent the policy from recommending clinically unacceptable ECMO adjustments. These masks should forbid pump flow above patient- or circuit-specific safe limits, sweep gas flow above predefined thresholds, FiO_2 below room air, abrupt changes beyond permitted rates, and actions that worsen drainage insufficiency or circuit pressure alarms [1-3]. In addition, actions should be blocked when recent data indicate severe instability, active bleeding, suspected oxygenator thrombosis, profound hypoxemia, or rapid vasopressor escalation. This safety layer operationalizes the principle that RL recommendations must remain subordinate to bedside safety constraints and established ECMO practice standards [4, 14, 15].

Lagrangian constraint formulation

Beyond hard masks, the framework can use constrained policy optimization in which the objective maximizes expected clinical reward while limiting expected safety cost. A Lagrangian formulation can penalize predicted complication probability, excessive oxygenator stress, high circuit pressure, unstable perfusion, or large deviations from clinician-supported actions, thereby embedding safety as a formal optimization constraint [10, 14, 16]. This is especially important in ECMO because the safest action may not be the action with the highest short-term gas exchange reward. A constrained policy can therefore recommend modest, behaviorally supported adjustments while avoiding actions that appear beneficial in value estimation but violate acceptable risk thresholds [1-3].

Real-Time Integration

Data pipeline

The real-time pipeline should connect bedside monitors, blood gas analyzers, ECMO consoles, ventilators, infusion systems, and electronic health record data into a synchronized feature stream. Data preprocessing should include timestamp alignment, artifact detection, unit harmonization, missingness handling, trend extraction, and uncertainty labeling before the state is passed to the RL inference engine [3-5]. The system should then generate a recommendation within a clinically useful time window while preserving an auditable record of input variables, excluded variables, action scores, safety-mask decisions, and final displayed recommendation. Similar clinical decision-support work in ECMO and critical-care RL emphasizes that model performance alone is insufficient without reliable integration into bedside workflow [4, 10, 14, 15].

Human-in-the-loop

The proposed system should present recommendations to clinicians as actionable suggestions, such as increasing pump flow by a small increment, reducing sweep gas flow, or maintaining current settings with reassessment after the next blood gas. Clinicians should be able to accept, modify, or reject each recommendation, and the interface should display the physiologic rationale, safety checks, recent trends, and uncertainty estimate [1-3]. This human-in-the-loop structure is necessary because ECMO management depends on context that may not be fully captured in structured data, including cannula position, bedside examination, bleeding concerns, ventilator synchrony, and goals of care. It also preserves clinical accountability while allowing the RL system to provide consistent, data-driven support [4, 14, 15].

Evaluation Strategy

Simulator validation

Initial evaluation should use historical ECMO trajectories to create a retrospective simulator for policy testing, counterfactual estimation, and failure-mode analysis. Off-

policy evaluation methods such as importance sampling, weighted importance sampling, fitted Q evaluation, and doubly robust estimation can estimate whether the learned policy would have selected actions associated with better long-term outcomes under observed data limitations [10, 14, 16]. However, ECMO simulators cannot fully resolve unmeasured confounding, clinician intent, or physiologic changes that were not recorded, so simulator findings should be treated as preclinical evidence rather than proof of benefit. Lessons from sepsis, ventilation, vasopressor, and sedation RL show that rigorous off-policy evaluation is necessary before prospective decision-support deployment [7, 8, 11, 12, 17].

Performance metrics

Performance metrics should include time to target oxygenation, time to target $PaCO_2$, pH stabilization, cumulative duration of extreme values, frequency of recommended adjustments, estimated complication rate, safety-violation rate, and reward per episode. ECMO-specific endpoints should also include successful decannulation, failure to wean, oxygenator exchange, circuit thrombosis, bleeding, hemolysis, neurologic injury, and mortality, because optimization of device settings is only meaningful if it improves patient-centered and circuit-safety outcomes [6, 18-22]. Evaluation should also measure concordance with clinician decisions and identify cases where the model recommends a different action, since these disagreements are the most informative for expert review. A clinically credible framework must therefore evaluate physiologic control, safety, interpretability, and alignment with expert practice rather than reporting a single aggregate reward value [3, 4].

Table 2 consolidates the safety, validation, and governance requirements that distinguish a clinically credible offline ECMO reinforcement learning framework from an unconstrained optimization model.

Table 2. Safety and Validation Requirements for a Clinically Credible Offline ECMO Reinforcement Learning System

Requirement layer	Core requirement	Failure mode addressed	M va exp
Dataset construction	Patient-level episode splitting, time-aligned device and physiologic data, complete outcome capture	Data leakage, irregular sampling bias, incomplete state representation	De rep prep mis h
Conservative policy learning	Penalize unsupported or out-of-distribution actions	Extrapolation beyond observed ECMO practice	C cor a cor po si
Reward specification	Balance gas exchange, hemodynamics, weaning, and harm avoidance	Blood-gas optimization at the expense of complications	S ar we p
Safety shield	Apply hard masks, action bounds, rate limits, and alarm-based exclusions	Unsafe recommendations reaching clinicians	b fr an vio
Off-policy evaluation	Use multiple retrospective evaluation methods rather than a single reward score	Overconfident claims from biased retrospective estimation	Im sa t ev dou es ex
Human-in-the-loop deployment	Present recommendations as advisory, explainable, and rejectable	Automation bias, unclear accountability, workflow disruption	Sil t u t c cor a
External validation	Test across centers, ECMO modes, devices,	Poor generalization across institutions	Mu va

	and patient subgroups	or cannulation strategies	pr

Baseline comparisons

The RL policy should be compared with historical clinician decisions, rule-based ECMO protocols, proportional-integral-derivative control concepts, and non-conservative RL variants that illustrate the dangers of unconstrained optimization. Historical clinician behavior provides the most realistic baseline because the offline dataset is generated by expert bedside practice, while rule-based protocols provide transparent comparators for gas exchange and weaning decisions [1, 2, 4]. Standard online RL is not an ethically appropriate bedside comparator for ECMO, but it can be included in simulation to show how unconstrained exploration differs from conservative offline learning [10, 14, 16]. Baseline comparisons should focus on whether the proposed framework improves consistency, reduces avoidable adjustment burden, and preserves safety margins without claiming clinical superiority before prospective validation [3, 15].

Limitations

Technical limitations

The main technical limitations are off-policy evaluation bias, incomplete observability, irregular sampling, unmeasured confounding, and the sim-to-real gap. ECMO decisions are affected by anticoagulation, transfusion, ventilator strategy, sedation, mobilization, cannula position, clinician judgment, institutional protocols, and goals of care, many of which may be inconsistently represented in structured datasets [1, 2, 4]. Offline RL can also overestimate the value of rarely observed actions or learn correlations that reflect treatment selection rather than causal benefit, even when conservative algorithms are used [10, 14, 16]. These limitations mean that the framework should be viewed as a research architecture requiring extensive audit, external validation, and prospective silent-mode testing.

Clinical limitations

Clinical limitations include clinician acceptance, medico-legal responsibility, alarm fatigue, workflow disruption, and uncertainty about how recommendations should be documented and governed. ECMO is a multidisciplinary therapy involving intensivists, perfusionists, nurses, surgeons, respiratory therapists, and transplant or heart-failure teams, so any RL tool must fit team-based decision-making rather than provide isolated algorithmic instructions [1, 2, 5]. A model trained on one institution, device platform, population, or cannulation strategy may not generalize to another without recalibration and external validation [6, 18, 21, 22]. For these reasons, the framework should be implemented first as advisory decision support under clinician control, with prospective validation required before any claim of improved outcomes [3, 4].

Conclusion

This manuscript presented a conceptual offline reinforcement learning framework for dynamic ECMO optimization using real-time blood gas, hemodynamic, pump flow, and circuit pressure data. The framework treats ECMO management as a sequential decision problem in which each setting adjustment changes the future physiologic and device state. It is designed to recommend incremental changes in pump flow, sweep gas flow, and FiO_2 rather than replace bedside clinicians. Its purpose is to define a safe research pathway for future development.

The key advantages of this approach are personalization, proactive support, offline training, and explicit safety constraint handling. A patient-specific policy can account for trends in gas exchange, perfusion, circuit behavior, and recovery rather than relying only on static thresholds. Offline learning avoids dangerous bedside exploration, while safety masks and constrained optimization reduce the chance of unacceptable recommendations. Human-in-the-loop review preserves clinical judgment and accountability.

The framework also has important limitations. It depends on historical data quality, accurate time alignment, sufficient representation of clinically relevant actions, and valid off-policy evaluation. Simulation and retrospective analysis cannot prove bedside benefit because ECMO physiology is complex and many determinants of clinical decisions are not fully captured in structured data. Prospective validation, silent deployment, usability testing, and governance review are required before clinical implementation.

Future work should develop this framework using large ECMO registries and multicenter datasets that capture device settings, blood gases, hemodynamics, circuit pressures, complications, and outcomes. Collaboration among intensivists, perfusionists, data scientists,

biomedical engineers, ethicists, and regulatory experts will be essential. The goal should not be autonomous ECMO control at the outset, but trustworthy decision support that improves consistency, safety, and personalization. With careful validation, offline reinforcement learning may become a foundation for real-time ECMO optimization in high-acuity care.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Sep 2025 Revised: 02 Dec 2025 Accepted: 08 Feb 2026
Published online: 20 July 2026

References

Tonna JE, Abrams D, Brodie D, Greenwood JC, Mateo-Sidron JA, Usman A, et al. Management of adult patients supported with venovenous extracorporeal membrane oxygenation (VV ECMO): guideline from the extracorporeal life support organization (ELSO). ASAIO J. 2021;67(6):601-10. https://doi.org/10.1097/MAT.0000000000001432.

Lorusso R, Shekar K, MacLaren G, Schmidt M, Pellegrino V, Meyns B, et al. ELSO interim guidelines for venoarterial extracorporeal membrane oxygenation in adult cardiac patients. ASAIO J. 2021;67(8):827-44. https://doi.org/10.1097/MAT.0000000000001499.

Song J, Dave SB, Yang Y, Foote H, Moore R, Upadhyaya P, et al. rECMOmender: reinforcement learning for decision support in venovenous extracorporeal membrane oxygenation management. Crit Care Explor. 2026;8(2):e1369.

Pladet L, Luijken K, Fresiello L, Miranda DD, Hermens JA, Smeden MV, et al. Clinical decision support for extracorporeal membrane oxygenation: will we fly by wire? Perfusion. 2023;38(1 Suppl):68-81. https://doi.org/10.1177/02676591221113227.

Cohen W, Mirzai S, Li Z, Combs P, Hu K, Rose R, et al. Personalized ECMO: crafting individualized support. J Cardiothorac Vasc Anesth. 2022;36(5):1477-86. https://doi.org/10.1053/j.jvca.2021.11.034.

Ayers B, Wood K, Gosev I, Prasad S. Predicting survival after extracorporeal membrane oxygenation by using machine learning. Ann Thorac Surg. 2020;110(4):1193-200. https://doi.org/10.1016/j.athoracsur.2020.02.006.

Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med. 2021;4(1):32. https://doi.org/10.1038/s41746-021-00388-6.

den Hengst F, Otten M, Elbers P, van Harmelen F, François-Lavet V, Hoogendoorn M. Guideline-informed reinforcement learning for mechanical ventilation in critical care. Artif Intell Med. 2024;147:102742. https://doi.org/10.1016/j.artmed.2023.102742.

Liu S, Xu Q, Xu Z, Liu Z, Sun X, Xie G, et al. Reinforcement learning to optimize ventilator settings for patients on invasive mechanical ventilation: retrospective study. J Med Internet

Res. 2024;26:e44494. https://doi.org/10.2196/44494.

Roggeveen LF, Hassouni AE, de Grooth HJ, Girbes AR, Hoogendoorn M, Elbers PW, et al. Reinforcement learning for intensive care medicine: actionable clinical insights from novel approaches to reward shaping and off-policy model evaluation. Intensive Care Med Exp. 2024;12(1):32. https://doi.org/10.1186/s40635-024-00625-5.

Kalimouttou A, Kennedy JN, Feng J, Singh H, Saria S, Angus DC, et al. Optimal vasopressin initiation in septic shock: the OVISS reinforcement learning study. JAMA. 2025;333(19):1688-98. https://doi.org/10.1001/jama.2025.4738.

Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716-20. https://doi.org/10.1038/s41591-018-0213-5.

Bologheanu R, Kapral L, Laxar D, Maleczek M, Dibiasi C, Zeiner S, et al. Development of a reinforcement learning algorithm to optimize corticosteroid therapy in critically ill patients with sepsis. J Clin Med. 2023;12(4):1513. https://doi.org/10.3390/jcm12041513.

Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: comprehensive review. J Med Internet Res. 2020;22(7):e18477. https://doi.org/10.2196/18477.

Jayaraman P, Desman J, Sabounchi M, Nadkarni GN, Sakhuja A. A primer on reinforcement learning in medicine for clinicians. NPJ Digit Med. 2024;7(1):337. https://doi.org/10.1038/s41746-024-01269-4.

Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv. 2021;55(1):1-36. https://doi.org/10.1145/3477600.

Lee HY, Chung S, Hyeon D, Yang HL, Lee HC, Ryu HG, et al. Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients. NPJ Digit Med. 2024;7(1):325. https://doi.org/10.1038/s41746-024-01340-0.

Hsu JC, Pai CH, Lin LY, Wang CH, Wei LY, Chen JW, et al. Machine learning-based first-day mortality prediction for venoarterial extracorporeal membrane oxygenation: the novel RESCUE-24 score. ASAIO J. 2026;72(2):117-28.

Kalra A, Bachina P, Shou BL, Hwang J, Barshay M, Kulkarni S, et al. Using machine learning to predict neurologic injury in venovenous extracorporeal membrane oxygenation recipients: an ELSO registry analysis. JTCVS Open. 2024;21:140-67. https://doi.org/10.1016/j.xjon.2024.07.015.

Kalra A, Bachina P, Shou BL, Hwang J, Barshay M, Kulkarni S, et al. Acute brain injury risk prediction models in venoarterial extracorporeal membrane oxygenation patients with tree-based machine learning: an extracorporeal life support organization registry analysis. JTCVS Open. 2024;20:64-88. https://doi.org/10.1016/j.xjon.2024.05.009.

Gao H, Huang X, Zhou K, Ling Y, Chen Y, Mou C, et al. Development and validation of a machine learning model for predicting mortality risk in veno-arterial extracorporeal membrane oxygenation patients. Sci Rep. 2025;15(1):41581. https://doi.org/10.1038/s41598-025-41581-0.

Leng A, Bachina P, Liu O, Shou B, Racz C, Giliver DA, et al. Enhancing survival prediction after venoarterial extracorporeal membrane oxygenation using machine learning. ASAIO J. 2025;71(6):10-97.

Chen S, Qiu X, Tan X, Fang Z, Jin Y. A model-based hybrid soft actor-critic deep reinforcement learning algorithm for optimal ventilator settings. Inf Sci. 2022;611:47-64. https://doi.org/10.1016/j.ins.2022.08.032.

Lee H, Yoon HK, Kim J, Park JS, Koo CH, Won D, et al. Development and validation of a reinforcement learning model for ventilation control during emergence from general anesthesia. NPJ Digit Med. 2023;6(1):145. https://doi.org/10.1038/s41746-023-00891-1.

Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak. 2019;19(Suppl 2):57. https://doi.org/10.1186/s12911-019-0795-0.

Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Mak. 2020;20(Suppl 3):124. https://doi.org/10.1186/s12911-020-01145-y.

Drudi C, Mollura M, Li-wei HL, Barbieri R. A reinforcement learning model for optimal treatment strategies in intensive care: assessment of the role of cardiorespiratory features. IEEE Open J Eng Med Biol. 2024;5:806-15. https://doi.org/10.1109/OJEMB.2024.3366845.

Wu X, Li R, He Z, Yu T, Cheng C. A value-based deep reinforcement learning model with human expertise in optimal

treatment of sepsis. NPJ Digit Med. 2023;6(1):15. https://doi.org/10.1038/s41746-023-00773-6.

Zhang T, Qu Y, Wang D, Zhong M, Cheng Y, Zhang M. Optimizing sepsis treatment strategies via a reinforcement learning model. Biomed Eng Lett. 2024;14(2):279-89. https://doi.org/10.1007/s13534-023-00327-6.

Choi Y, Oh S, Huh JW, Joo HT, Lee H, You W, et al. Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. Commun Med (Lond). 2024;4(1):245. https://doi.org/10.1038/s43856-024-00662-8.