Deep Reinforcement Learning for Efficient and Fair Allocation of Healthcare Resources

Yikuan Li[1,*], **Chengsheng Mao**[1,*], **Kaixuan Huang**[2,*], **Hanyin Wang**[1,*], **Zheng Yu**[2,*], **Mengdi Wang**[2,†] and **Yuan Luo**[1,†],
[1]Northwestern University
[2]Princeton University
{yikuan.li, chengsheng.mao, hanyin.wang, yuan.luo}@northwestern.edu,
{kaixuanh, zhengy, mengdiw}@princeton.edu

Abstract

The scarcity of health care resources, such as ventilators, often leads to the unavoidable consequence of rationing, particularly during public health emergencies or in resource-constrained settings like pandemics. The absence of a universally accepted standard for resource allocation protocols results in governments relying on varying criteria and heuristic-based approaches, often yielding suboptimal and inequitable outcomes. This study addresses the societal challenge of fair and effective critical care resource allocation by leveraging deep reinforcement learning to optimize policy decisions. We propose a transformer-based deep Q-network that integrates individual patient disease progression and interaction effects among patients to enhance allocation decisions. Our method aims to improve both fairness and overall patient outcomes. Experiments using metrics such as normalized survival rates and interracial allocation rate differences demonstrate that our approach significantly reduces excess deaths and achieves more equitable resource allocation compared to severity- and comorbidity-based protocols currently in use. Our findings highlight the potential of deep reinforcement learning to address critical health care challenges.

1 Introduction

The Institute of Medicine (IOM) defines Crisis Standards of Care as "a substantial change in usual health care operations and the level of care it is possible to deliver, which is made necessary by a pervasive (e.g., pandemic influenza) or catastrophic (e.g., earthquake) disaster" [Gostin *et al.*, 2012]. These guidelines recognize that pandemics can strain health systems into an absolute scarcity of health care resources and could result in the unavoidable consequence of rationing.

Following the IOM framework, state governments throughout the U.S. have developed allocation protocols for critical care resources during the pandemic [Piscitello *et al.*, 2020]. Consistent with the broad consensus of ethicists and stakeholders [Emanuel *et al.*, 2020], these protocols aim to triage patients via a pre-specified and transparent policy.

***Societal Challenges**:* Despite adhering to a general framework, critical details of these protocols vary widely across the U.S. [Piscitello *et al.*, 2020]. The protocols also differ in whether they prioritize younger patients or those without pre-existing medical conditions. For example, the SOFA protocol (used in New York [VEN, 2015]) focuses on maximizing short-term survival and does not consider age or pre-existing conditions; multiprinciple protocols (used in Maryland and Pennsylvania) give preference to younger patients and those without comorbidities [Biddison *et al.*, 2019]. However, all these protocols assign a static priority score, lacking the flexibility to account for the severity of scarcity and the specific patient pool they are competing against, which leaves room for improvement in patient outcomes.

Beyond the lack of a universally accepted protocol, there is also a pressing societal challenge regarding health equity in resource allocation. Empirical assessments indicate that existing allocation protocols disproportionately disadvantaged Black patients, who were significantly less likely to receive ventilators [Bhavani *et al.*, 2021]. Furthermore, the SOFA score, widely used in triage decisions, has been shown to be only moderately effective in predicting mortality for mechanically ventilated patients [Raschke *et al.*, 2021]. Additionally, evidence suggests that SOFA scores may inadvertently assign higher severity scores to Black patients, which could lead to their lower prioritization for life-saving interventions. [Ashana *et al.*, 2021] These disparities highlight the urgent need to improve both the effectiveness and fairness of allocation protocols to ensure that critical resources are allocated efficiently and equitably among all patients.

Multidisciplinary Collaboration Given that resource allocation decisions are made sequentially and must adapt to evolving patient conditions, reinforcement learning (RL) provides a natural framework for optimizing allocation protocols. RL can leverage large-scale patient data and has the potential to learn optimal allocation strategies that maximize lives saved while ensuring fairness across demographic groups. Developing such an approach requires a multidisciplinary effort that brings together health informaticians and AI researchers. By integrating expertise in clinical decision-making, health data science, and AI, we can design an RL

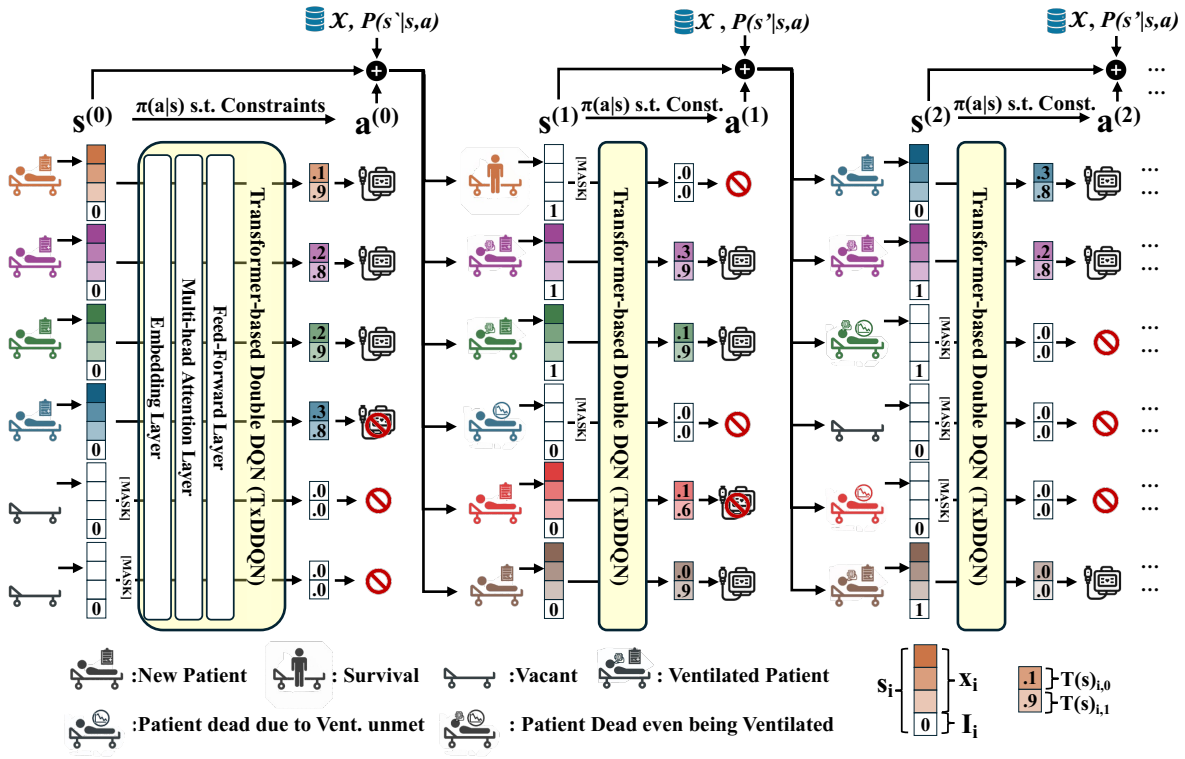

Figure 1: An illustration for our study formulation. Each color represents a separate patient.

model tailored to the complexities of real-world health crises.

In this study, we investigate the use of RL for optimizing critical care resource allocation policies. Our goal is to develop an adaptive, data-driven allocation policy that maximizes survival rates while promoting fairness in rationing across different racial populations. Our main contributions are summarized as follows:

- We formulate fair health care resource allocation as a multi-objective deep reinforcement learning problem, by integrating the utilitarian and egalitarian objectives into the RL rewards.
- We propose a Transformer-based parametrization of the deep Q-network that significantly reduces the complexity of the classical deep-Q network while making allocation decisions based on individual patient disease progression and interactions among patients.
- We apply our approach to a large, diverse, multi-hospital real-world clinical datasets. Experiments show that our approach leads to fair allocation of critical care resources among different races, while maintaining the overall utility with respect to patient survival.

2 Related Work

Health care resource scarcity, especially during pandemics and in intensive care settings, requires decision strategies that balance outcomes, equity, and crisis management [Emanuel and Wertheimer, 2006; Truog *et al.*, 2006; Persad *et al.*, 2009]. Despite these goals, implicit and explicit discrimination has long permeated healthcare, resulting in numerous instances of biased outcomes [Dresser, 1992; Tamayo-Sarver *et al.*, 2003; Chen *et al.*, 2008].

RL, given its inherent capacity for goal-oriented learning and sequential decision-making, holds promise for developing optimized allocation strategies. However, like other machine learning methods, RL can inadvertently reinforce disparities by favoring majority groups and failing to incorporate fairness objectives [Mehrabi *et al.*, 2021]. Indeed, RL may further exacerbate inequities across patient populations if fairness considerations are neglected [Liu *et al.*, 2018; Ahmad *et al.*, 2020; Rajkomar *et al.*, 2018; Pfohl *et al.*, 2021; Wang *et al.*, 2022; Li *et al.*, 2022].

Existing work on resource allocation has primarily focused on one-time or rollout-based distribution, such as vaccine allocation [Awasthi *et al.*, 2022; Rey *et al.*, 2023; Cimpean *et al.*, 2023], often using multi-armed bandit approaches. These methods, however, are ill-suited for complex scenarios requiring repeated, daily allocation decisions.

RL has also been applied to simulate pandemic trajectories, guide lockdown strategies [Zong and Luo, 2022], project ventilator needs [Bednarski *et al.*, 2021], and allocate PCR tests for COVID-19 screening [Bastani *et al.*, 2021]. For a comprehensive summary of RL applications in healthcare, see the survey by [Yu *et al.*, 2021]. Nonetheless, many of these efforts prioritize efficiency over long-term fairness, leaving critical gaps in addressing disparities in high-stakes resource allocation.

3 Problem Formulation

We formulate the ventilator allocation problem as a day-to-day sequential decision problem, as illustrated in Figure 1. **Our objective is to optimize the triage protocol regarding who should receive health resources under scarcity, with the goal of saving more lives during a health crisis. This optimization by no means influences physicians' medical decisions regarding whether a patient should be ventilated or not.** Therefore, in our formulation and experiments, all patients are prescribed ventilation by physicians but might not receive one due to resource limitation.

3.1 State Space

The state space $\mathcal{S}$ describes the current clinical conditions of all patients in the hospital, as well as their ventilation status. Each state is represented by $s = [x_1, x_2, \cdots, x_N, I_1, I_2, \cdots, I_N] \in \mathcal{S} \subseteq \mathbb{R}^{kN+N}$, where $x_i \in [0,1]^k \subseteq \mathbb{R}^k$ denotes the current medical condition of the patient on bed i, and indicator $I_i \in \{0,1\}$ denotes whether the bed i has been ventilated. Apart from normal medical conditions of patients, we consider three special conditions: `Survived`, `Dead`, and `Vacant`, corresponding to the cases where the patient in this bed is recovered or dead after ventilation, as well as currently no patient in this bed. They act as terminators for patients or separators between patients in the same bed. Such designs separate different patients on the same bed explicitly, so that the tasks of learning the progression of medical conditions given ventilation and the task of recognizing the end of each patient can be decoupled. The intensive care units have up to N beds. Here we only consider ventilator scarcity, but bed scarcity can be analogously modeled. In the following, we give a rigorous formulation of the RL model for both cases of without and with the consideration of the fairness in distributing ventilators.

With fairness in consideration, we further record the cumulative numbers of total and ventilated patients of different ethnoracial groups in the state vector, denoted by $n_k, m_k \in \mathbb{R}$ respectively, where $k \in \{B, W, A, H\}$ denotes 4 ethnoracial group in the dataset: non-Hispanic Black, non-Hispanic White, non-Hispanic Asian, Hispanic. Thus, each state $s = [x_1, \cdots, x_N, I_1, \cdots, I_N, n_B, n_W, n_A, n_H, m_B, m_W, m_A, m_H] \in \mathcal{S} \subseteq \mathbb{R}^{kN+N+8}$ describes the medical and ventilation status of all current patients, as well as the number of cumulative total and ventilated patients of different groups.

3.2 Action Space

For both cases, the action space $\mathcal{A}$ is a discrete space denoting whether each bed is on ventilation or not. Let us use 1 to denote ventilate and 0 otherwise, so the action space $\mathcal{A} \subseteq \{0,1\}^N$. Note that we have two constraints on actions:

- Capacity Constraint: $\mathcal{A} \subseteq \{a \in \{0,1\}^N : \sum_{i=1}^N a_i \leq C\}$, where C is the ventilator capacity of the hospital.
- Withdrawal Constraint (optional): a patient who has been ventilated must not be withdrawn until they no longer need it or are discharged.: $\mathcal{A} \subseteq \{a \in \{0,1\}^N : a_i = 1 \text{ if } I_i = 1, i = 1, 2, \cdots, N\}$, where I_i in the state information denotes whether current patient on bed i has been ventilated before. We apply this constraint following [Bhavani *et al.*, 2021]. However, in real clinical practice, the withdrawal of a ventilator from one patient to save another raises ethical issues and has not reached a consensus (e.g., ranging from no mechanism in the Maryland protocol to an explicit SOFA-based approach in the New York protocol). Therefore, we also include the results without this constraint in Appendix A.5 [1] to provide a comprehensive picture across the spectrum, considering both extremes where withdrawal is either considered or not at all.

Therefore, the action space is shrunk to $\mathcal{A} = \{a \in \{0,1\}^N : \sum_{i=1}^N a_i \leq C \text{ and } a_i = 1 \text{ if } I_i = 1, \forall i = 1, 2, \cdots, N\}$.

3.3 Transition Model

For ease of notation, we denote the three special conditions `Survived`, `Dead`, and `Vacant` as $\mathbf{1}, -\mathbf{1}, \mathbf{0} \in \mathbb{R}^k$ respectively. In the case without fairness consideration, given current state $s = [x_1, x_2, \cdots, x_N, I_1, I_2, \cdots, I_N]$ and action $a = [a_1, a_2, \cdots, a_N]$, it will transit to the next state $s' = [x'_1, x'_2, \cdots, x'_N, I'_1, I'_2, \cdots, I'_N]$ in the following coordinate-wise way:

- If $x_i \neq \mathbf{0}, \mathbf{1}, -\mathbf{1}$, then the patient will transit to `Dead` condition if not ventilated:

$$P_i^{\text{clinical}}(x'_i|s,a) = \begin{cases} 1 & \text{if } a_i = 0, x'_i = -\mathbf{1}, \\ 0 & \text{if } a_i = 0, x'_i \neq -\mathbf{1}, \\ p^{\text{on}}(x'_i|x_i) & \text{if } a_i = 1, x'_i \neq \mathbf{0}, \\ 0 & \text{if } a_i = 1, x'_i = \mathbf{0}, \end{cases}$$

  where $p^{\text{on}}(x'_i|x_i)$ denotes the probability of a patient transiting from medical condition x_i to x'_i given ventilation. The ventilation status is naively transited $P_i^{\text{vent}}(I'_i|s,a) = 1$ if $I'_i = a_i$ and 0 otherwise. We did not use computational methods to simulate $p^{\text{on}}(x'_i|x_i)$, as the progression of patients' conditions is high-dimensional and difficult to model. Instead, we only used real clinical trajectories by sampling from real-world clinical databases.

  Following existing clinical literature (e.g., [Bhavani *et al.*, 2021]), we assume that patients who needed ventilators but did not receive one will die. Thus, there will be no patients *waiting* for a ventilator, as the inability to receive ventilation will result in their immediate deceased and removal from the dataset. However, such an assumption might be over-pessimistic. Therefore, in Appendix A.6, we also explore scenarios where patients not being allocated ventilators does not lead to immediate fatalities, which validates the broader applicability of our proposed methods.

- If $x_i = \mathbf{0}, \mathbf{1}$ or $-\mathbf{1}$, the action does not influence the transition dynamic:

$$P_i^{\text{clinical}}(x'_i|s,a) = \begin{cases} 0 & \text{if } x'_i = \mathbf{1} \text{ or } -\mathbf{1}, \\ 1 - q_i(s) & \text{if } x'_i = \mathbf{0}, \\ q_i(s) \cdot \xi(x'_i) & \text{if } x'_i \neq \mathbf{0}, \mathbf{1}, -\mathbf{1}, \end{cases}$$

[1] https://arxiv.org/pdf/2309.08560

where $\xi(\cdot)$ denotes the distribution of the initial medical condition of a patient when admitted to the critical care units, and $q_i(s)$ denotes the probability of a new incoming patient staying in bed i. Note $q_i(s)$ depends on how new patients are distributed to empty beds. It satisfies $q_i(s) = 0$ if the bed i is already occupied, and $\sum_{i=1}^{N} q_i(s) = E \sim \text{Poisson}(\Lambda)$ assuming the number of incoming patients E obeys an Poisson distribution with parameter Λ. The ventilation status does not matter at this point, we can set $P_i^{\text{vent}}(I_i'|s,a) = 1$ if $I_i' = 0$ and 0 otherwise.

Given above discussion for transition dynamics of each individual patients/bed, the overall transition probability can be written as

$$P = (P_0^{\text{clinical}}, \cdots, P_N^{\text{clinical}}, P_0^{\text{vent}}, \cdots, P_N^{\text{vent}}) \qquad (1)$$

We further consider the progress of the number of cumulative patients and ventilated patients of each ethnic group, whose deterministic transitions are naive by their definitions.

3.4 Reward

The reward function consists of the following three parts:

- Terminal condition $\mathbf{R_t}$: If a patient is discharged alive after being on a ventilator, a positive reward of 1 is given. If a patient requires a ventilator but is not able to receive one, or dies after being on a ventilator, a penalty of -1 is given: $R_t(s,a) = \sum_{i=1}^{N} \mathbf{1}[x_i = \mathbf{1}] - \sum_{i=1}^{N} \mathbf{1}[x_i = -\mathbf{1}]$.
- Ventilation cost $\mathbf{R_v}$: if a ventilator is used, it occurs a small negative reward:$R_v(s,a) = \sum_{i=1}^{N} a_i$.
- Fairness penalty $\mathbf{R_f}$: in the case with fairness consideration, we consider the cumulative total and ventilated patients of different ethnoracial groups. We expect the distribution of ventilators is equitable in terms of the proportion of ventilated patients of all ethnoracial groups. Therefore, a penalty of KL-divergence between the frequency distributions of incoming patients of different ethnoracial groups $\mathcal{D}_n \overset{d}{\sim} [n_B, n_W, n_A, n_H]/(n_B + n_W + n_A + n_H)$ and ventilated patients of different ethnoracial groups $\mathcal{D}_m \overset{d}{\sim} [m_B, m_W, m_A, m_H]/(m_B + m_W + m_A + m_H)$ is considered: $R_f(s,a) = \text{KL}(\mathcal{D}_n \| \mathcal{D}_m)$.

Thus, the reward function is given by

$$R(s,a) = R_t(s,a) + \mu \cdot R_v(s,a) + \lambda \cdot R_f(s,a), \qquad (2)$$

where the parameter $\lambda \geq 0$ balances the trade-off between ventilation effectiveness and fairness. In the case without fairness consideration, we set $\lambda = 0$. μ is a small scalar that controls ventilation costs and can be selected through parameter fine-tuning or clinical expertise.

4 Method

In our formulation, due to the resource constraints imposed through the restricted action sets and the fairness requirements via reward penalties, the interaction effects among the patients need to be considered. However, modeling all patients jointly poses computational and memory challenges due to the combinatorial nature of the action space. Addressing such a formulation necessitates a Q network with dimensions proportional to the size of the state space $|\mathcal{S}|$ and action space $|\mathcal{A}|$, resulting in a complexity up to $O(N \times 2^N)$, where N represents the ICU bed capacity. This complexity makes it impractical to naively construct and train such a Q network, given the challenges in collecting a sufficient number of data points. In practical terms, the intensive care units of health systems typically have hundreds of beds ($N \geq 100$), making the computational demands and data requirements for training such a network unattainable. Also, naively concatenating all patients' state vectors introduces an order among patients, which may cause the model to learn artificial factors on that order and potentially bias the model.

To circumvent the computational intractability without losing the consideration of the interaction effects among patients, we therefore propose Transformer-based Q-network parametrization, which inherits the classical Q-learning framework with new parametrization and greedy action selection tailored to our problem structure. In our formulation, the transition of states is partially decomposable because a single patient's clinical conditions depend only on their preceding clinical conditions and ventilation status. Therefore, we can reshape the state from a one-dimensional long vector to a two-dimensional matrix, with the i-th row as $[x_i, I_i]$ for patient i. In case the fairness features n_k, m_k are also considered, we can replicate them N times and append them to the end of each row. The Q-network is parametrized as $T_\theta : \mathbb{R}^{\dim(\mathcal{S})} \to \mathbb{R}^{N \times 2}$, whose input is the reshaped state matrix. The i-th row of the output $T_\theta(s)_i \in \mathbb{R}^2$ corresponds to the Q-value contributed by the i-th patient given the current state s and the action a_i on the i-th patient. We adopt an additive form of the joint Q-value, which estimates the trade-off between effectiveness and fairness when allocating under the constraints and considering the clinical conditions of all patients: $Q(s,a) = \sum_{i=1}^{N} T_\theta(s)_{i,a_i}$. We leverage the Transformer architecture [Vaswani *et al.*, 2017] for T_θ. Each row in the reshaped s are considered as an input patient token. As our input to the model is essentially an orderless set of patients' conditions with indefinite size, Transformers are a natural fit for our problem because Transformers without positional encodings are permutationally-invariant and can handle inputs with arbitrary sizes. Each Transformer layer is composed of a feed-forward layer and a multi-head attention layer. The feed-forward layer acts independently on each element of the input as a powerful feature extractor, while the attention layer is able to capture the interaction effects between all the elements. This quadratic complexity of Transformer architecture also results in a significant reduction in the network's complexity, decreasing it from $O(N \times 2^N)$ to a more manageable $O(N^2)$. We also provide robust empirical results to demonstrate the effectiveness and advantages of our proposed transformer-based parametrization compared to the classical deep Q-network in Appendix A.4.

In this parametrization, the greedy action shall be searched within the valid action space $\mathcal{A}$ to accommodate the withdrawal and capacity constraints: $\pi_\theta(a|s) : a^* = \text{argmax}_{a \in \mathcal{A}} \sum_{i=1}^{N} T_\theta(s)_{i,a_i}$. Under our parametrization and

the additive form of the Q-value, this constrained greedy search can be efficiently solved by first allocating ventilators to those who have already been allocated (withdrawal constraint), where $a_i^* = 1 \Leftrightarrow I_i = 1$. Subsequently, the remaining $(C - \sum_{i=1}^{N} I_i)$ ventilators are allocated to the newly admitted patients for whom the ventilation improvement $d_i = T_\theta(s)_{i,1} - T_\theta(s)_{i,0}$ ranks at top-$(C - \sum_{i=1}^{N} I_i)$ among all their competitors (capacity constraints).

We also develop a simulator, $\text{Simu}(\mathcal{X}; C, \Lambda)$, constructed from real-world clinical trajectories to create the replay buffer for training the proposed model. This simulator maintains the clinical trajectory of each patient while randomizing their relative admission order. $\mathcal{X}$ is a set of clinical trajectories from all patients in the training cohort, with each trajectory being of various lengths covering all the clinical conditions of a patient in the critical care unit. At each time step in the simulator, we sample the initial conditions of E patients from $\mathcal{X}$ without replacement. E is determined by a Poisson distribution with parameter Λ, inferred from the distribution of ventilator requests in the training cohort. We apply a protocol Π to allocate the available C ventilators when the sum of existing patients and newly admitted E patients exceeds the capacity C. Patients who are not allocated ventilators or those who are discharged (either alive or deceased) are removed from the ongoing simulator and returned to the sampling pool. Patients who are allocated ventilators progress to the next time step of their own trajectory based on the transition function $p^{on}(x_i'|x_i)$. With this simulator setup, we can generate MDP tuples of indefinite length.

We combine the **Trans**former q-network parametrization with **D**ouble-**DQN** (**TxDDQN**) [Van Hasselt *et al.*, 2016], a variant to the original DQN [Mnih *et al.*, 2013] capable of to reduce overestimation. We summarize our proposed model and simulator in Algorithm 1.

Algorithm 1 Transformer-based Double Deep Q Network

Input: $\text{Simu}(\mathcal{X}; C, \Lambda)$, discount factor γ, learning rate α, batch size $|\mathcal{B}|$, loss function $\mathcal{L}_c$,network update frequency h, network update parameter τ
Initialize: Primary network T_{θ_0}; Target network $T_{\theta'}, \theta' \leftarrow \theta_0$; Allocation protocol $\Pi \leftarrow \pi_{\theta_0}$; $\mathcal{D} \leftarrow \Phi$; $s^{(0)}$ by sampling initial medical conditions of $\text{Pois}(\Lambda)$ patients from $\mathcal{X}$
for $e = 0$ **to** E ▷ *Training loop in epochs*
  for $t = 0$ **to** $T - 1$ ▷ *Construct a ring replay buffer*
    $a \sim \pi_{\theta_e}(s)$
      ▷ $\pi_\theta(a|s) : a^* = \text{argmax}_{a \in \mathcal{A}} \sum_{i=1}^{N} T_\theta(s)_{i,a_i}$
    $s' \sim P(s, a)$ ▷ *Eq.* (1) *for* $P(s, a)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s, a, s', R(s, a)\}$ ▷ *Eq.* (2) *for* $R(s, a)$
  $\theta_{e,0} \leftarrow \theta_e$
  for $g = 0$ **to** $G - 1$ ▷ *Gradient steps*
    Sample mini-batch $\mathcal{B} \subset \mathcal{D}$ ▷ $\mathcal{B} = \{(s, a, s', r\}^{|\mathcal{B}|}$
    for $i = 0$ **to** N ▷ *Search next action* a'^*
      $d_i = (T_{\theta_{e,g}}(s')_{i,1} - T_{\theta_{e,g}}(s')_{i,0})$ ▷ *Vent. improv.*
      if $I_i' = 1$ **then**
        $a_i' \leftarrow 1$ ▷ *Withdrawal constraint*
      elif d_i ranks top- $(C - |I'|)$ in $\{d_i \mid I_i' = 0\}$**then**
        $a_i' \leftarrow 1$ ▷ *Capacity constraint*
      else $a_i' \leftarrow 0$
    $\theta_{e,g+1} \leftarrow \theta_{e,g} - \alpha \cdot \nabla_\theta \mathcal{L}_c[\sum_{i=1}^{N} T_{\theta_{e,g}}(s)_{i,a_i}, R(s, a) + \gamma \cdot \sum_{i=1}^{N} T_{\theta'}(s')_{i,a_i'}]$ ▷ *Policy update*
    if $g \bmod h = 0$ **then** $\theta' \leftarrow \tau \cdot \theta_{e,g+1} + (1 - \tau) \cdot \theta'$
  $\theta_e \leftarrow \theta_{e,G}$

5 Experiments

5.1 Dataset

The dataset is sourced from Northwestern Medicine Enterprise Data Warehouse (NMEDW), including electronic health records, pathology data from multiple real-world hospitals and research laboratories. We collected 11,773 ICU admissions that have been allocated mechanical ventilators between March 15, 2020 and January 15, 2023. We filtered the patients with age between 18 and 95 for consideration. We removed admissions with a ventilator allocation duration exceeding 30 days to eliminate anomalies in the data. We extracted 38 features for each admission, including SOFA components, vital signs, demographics, comorbidities (see Appendix A.1 for the list of features and their statistics).

We split our data into 3 splits, 5,455 admissions between March 15, 2020 and July 14, 2021 were used as training data; 1,047 admissions between July 15, 2021 and October 14 2021 were used as validation data on which we selected the best hyper-parameters for testing; 5,271 admissions between October 15, 2021 and January 15, 2023 were used as test data. The patient distribution of different races and the ventilator demands on each day are shown in Appendix A.1. We assume that the original dataset obtained from health systems reflects a scenario where there was an abundant supply of ventilators available. We aim to investigate the effectiveness of various protocols in mitigating excess deaths in ventilator scarcity.

This setup differed from conventional RL settings, which are trained and evaluated solely on a simulator. We had separate validation and testing sets. This design allowed us to compare the proposed method with existing protocols in real-world hospital operational settings, enhancing the effectiveness and generalizability of our model. Additionally, ventilator demands may fluctuate due to seasonality and outbreaks. A test set spanning a whole year can assess our method's vulnerability to these seasonal request surges.

5.2 Baseline Protocols

Our proposed method was compared with the following existing triage protocols: **Lottery**: Ventilators are randomly assigned to patients who are in need. **Youngest First**: The highest priority is given to the youngest patients. **SOFA**: Patients' prioritization is discretized into three levels (0-7: high, 8-11: medium, and 11+: low) with the lottery serving as the tiebreaker. **Multiprinciple (MP)**: Each patient is assigned a priority point based on their SOFA score (0-8: 1, 9-11: 2, 12-14: 3, 14+: 4). Patients with severe comorbidities receive an additional 3 points. In case of ties, priority is given to patients in a younger age group (Age groups: 0–49, 50–69, 70–84, and 85+). If ties still exist, a lottery is conducted to determine the final allocation. **Decision Tree (DT)**: Grand-Clément et al. [Grand-Clément *et al.*, 2021] introduced a data-driven decision-tree-based approach for optimizing ven-

	Survival	**Fairness**	**Allocation Rates**				
	Survival, %	DPR, %	Overall, %	Asian, %	Black, %	Hispanic, %	White, %
Lottery	75.17 ±0.45	**96.89 ±1.55**	75.60 ±0.28	74.95 ±1.81	75.65 ±1.10	76.08 ±0.90	75.68 ±0.33
Youngest	77.24 ±0.07	86.39 ±0.27	75.65 ±0.09	75.59 ±0.50	81.73 ±0.31	84.44 ±0.30	72.94 ±0.12
SOFA	80.88 ±0.32	92.37 ±1.01	77.92 ±0.25	75.06 ±1.44	73.58 ±0.77	78.66 ±1.24	79.10 ±0.45
DT	76.16 ±0.01	94.35 ±0.01	75.81 ±0.01	79.59 ±0.00	75.14 ±0.01	77.34 ±0.00	75.58 ±0.00
MP	81.99 ±0.14	90.68 ±0.95	78.34 ±0.20	77.91 ±1.48	74.33 ±0.66	81.96 ±0.56	78.81 ±0.34
TxDDQN	84.76 ±0.24	86.91 ±3.45	81.80 ±0.26	78.03 ±0.99	72.48 ±0.74	81.71 ±0.44	84.45 ±0.22
TxDDQN-fair-off	85.29 ±0.18	95.27 ±1.42	**81.96 ±0.15**	80.50 ±2.23	80.07 ±1.98	81.44 ±1.05	82.70 ±0.34
TxDDQN-fair	**85.41 ±0.23**	95.24 ±1.65	81.90 ±0.24	80.01 ±1.79	79.95 ±1.05	81.26 ±1.24	82.80 ±0.42

Table 1: Impact of triage protocols on survival, fairness, and allocation rates with limited ventilators (B = 40) corresponding to about 50% scarcity. Fairness metric is demographic parity ratio (DPR, the ratio between the smallest and the largest allocation rate across patient groups, 100% indicating non-discriminative). Standard deviations are from 10 experiments with different seeds. We bold the protocols with the highest survival rate, DPR, and overall allocation rates. We underline the protocols that fall within one standard deviation of the best result.

tilator allocation. Their protocol factors in the BMI, age, and SOFA score, classifying patients into two priority levels. Admission time is the tie-breaker within the same priority level.

5.3 Off-policy and offline training

The maximum daily demand for ventilators in the validation and testing sets is 85. Therefore, we trained 85 capacity-specific protocol models $C = 1, 2, \cdots, 85$. Ventilator capacity is normalized from [0, 85] to [0%, 100%]. We used $\Lambda = 12$ because the number of daily newly admitted patients to the critical care units in our dataset follows a Poisson distribution with $\Lambda = 12$. Our simulator supports both off-policy and offline RL training settings [Levine *et al.*, 2020]. In the off-policy RL setting, we iteratively update Π with π_θ when constructing the replay buffer. We refer to the off-policy trained model without and with fairness consideration as **TxDDQN** and **TxDDQN-fair**. In the offline setting, we use existing heuristic-based protocol, MP, as Π to create a large training set and do not update the training data during the training process. This offline training process mimics early-stage health crisis operations where resources are allocated using existing protocols, and improvements are sought thereafter. We refer the offline trained model with fairness consideration as **TxDDQN-fair-off**. Our experiments were conducted on firewall-protected servers. The training time for each capacity-specific model was less than 30 minutes using a single GPU. For hyper-parameter selections and the survival-fairness Pareto frontier, please see Appendix A.2.

5.4 Evaluation

We evaluated the performance of our proposed protocol based on normalized survival rates, with the survival number at no shortage in ventilators as 100% and no patients alive when no ventilator is available as 0%. Fairness was quantified by comparing the allocation rates across four ethnoracial groups. The allocation rate was calculated by dividing the total number of ventilators allocated by the sum of the ventilators requested. The demographic parity ratio (DPR) [Bird *et al.*, 2020] served as the group metric of fairness, and it is defined as the ratio between the smallest and largest group-level allocation rate. A DPR value close to 1 signifies an equitable allocation, indicating a non-discriminatory distribution among the different groups. Additionally, to gain insights into the impact of protocols under varying shortage levels, we visualized the survival-capacity curve (SCC) and allocation-capacity curve (ACC). The Area Under the Survival-Capacity Curve (AUSCC) serves as an indicator of the overall performance in terms of life-saving abilities under different levels of ventilator shortage. Similarly, the Area Under the Allocation-Capacity Curve (AUACC) reflects the overall performance of ventilator utilization rates under varying levels of shortage.

5.5 Results

Our findings regarding the survival rates for different triage protocols under various levels of ventilator shortage are illustrated in Figure 2. Across all triage protocols, higher ventilator capacities are associated with increased allocation rates, thereby resulting in saving more lives. In the left Panel of Figure 2, our proposed TxDDQN models exhibit higher AUSCC compared to all other baseline protocols, demonstrating the superiority of our models in terms of life-saving efficacy. Likewise, in the right Panel, TxDDQN models also demonstrate higher AUACC compared to other baselines, indicating their superiority in terms of ventilator utilization.

Figure 3 showcases the allocation-capacity curves for different ethnoracial groups and triage protocols in Panels A through H. Only Lottery, TxDDQN-fair and TxDDQN-fair-off model exhibit minimal disparities, but TxDDQN-fair and TxDDQN-fair-off surpasses Lottery in terms of allocation rates and life-saving efficacy. Conversely, all other triage protocols display a preference for specific ethnoracial groups. For example, Youngest and MP favor Hispanic, while SOFA and TxDDQN favor White.

We conducted a detailed analysis in Table 1 under the scenario where approximately 50% of ventilators are unavailable. Our results show that TxDDQN-fair achieves the highest survival rate and allocation rate among the triage protocols. It ranks second in terms of DPR, following closely behind the Lottery protocol. These findings confirm the effectiveness of our TxDDQN-fair approach in improving both survival rates and fairness simultaneously. Importantly, the inclusion of fairness rewards does not compromise its life-saving capabilities compared to TxDDQN. We also did not

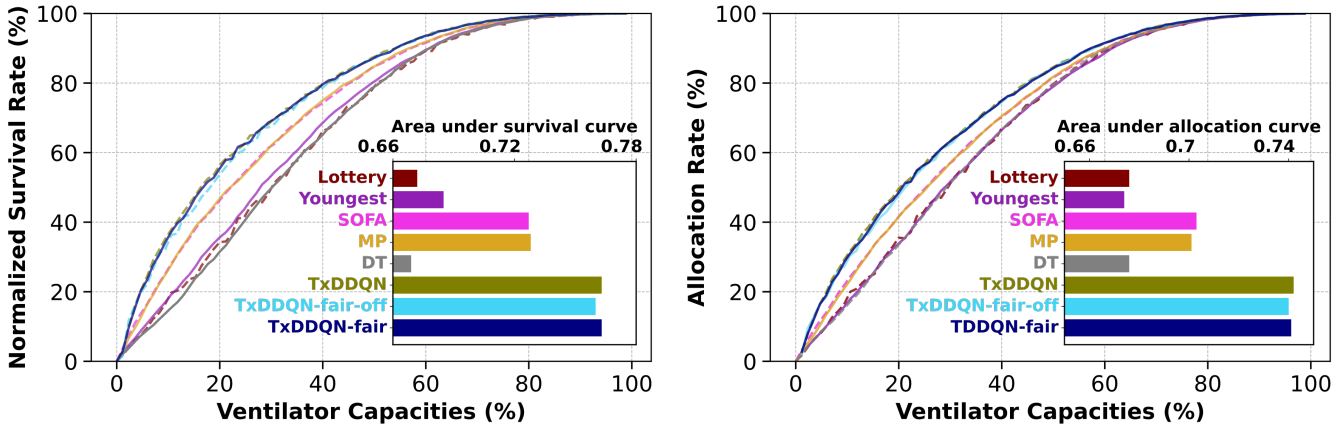

Figure 2: Impact of triage protocols on survival rates and allocation rates under varying levels of ventilator shortages. The bar plot associated with each panel indicates the area under the survival-capacity curve and allocation-capactiy curve, respectively, where a larger value indicates that the protocol can save more lives across different levels of shortages. Notably, the MP and SOFA curves exhibit overlap, indicating similar allocation patterns. Similarly, the lottery and youngest curves show close proximity, as do our three TxDDQN configurations.

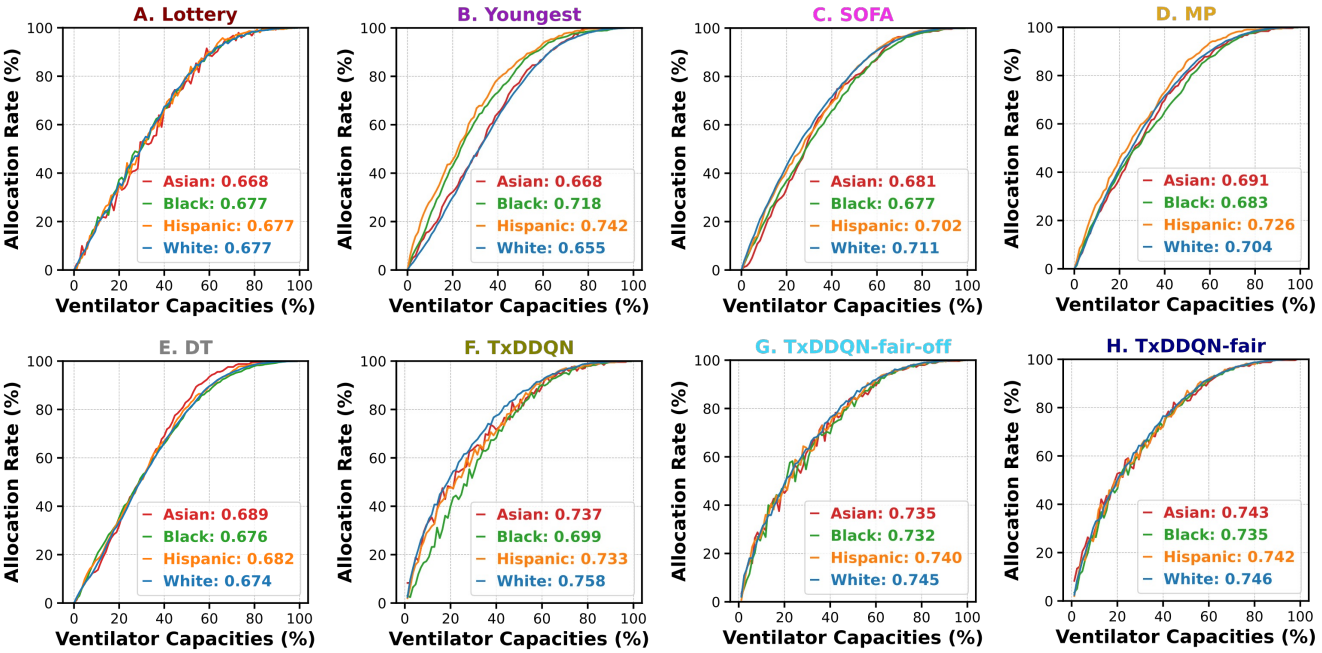

Figure 3: Allocation rates across protocols and ethnoracial groups. Each panel illustrates how allocation rates vary by ethnoracial group under different protocols. The numbers in the legend indicate the area under the allocation-capacity curve (AUACC).

observe differences between TxDDQN-fair and TxDDQN-fair-off in their life-saving abilities and fairness. This confirms that our proposed methods are adaptable to both offline and off-policy training, and provide the foundation for safe deployment in the early stages of health crises. Our proposed model also demonstrates enhanced fairness and life-saving outcomes in two additional settings: i) where withdrawal constraint is removed (Appendix A.5), and ii) the shortage of ventilators does not lead to immediate deaths (Appendix A.6). We provide ablation studies in Appendix A.3.

6 Summary

In this study, we formulated fair health care resource allocation as a multi-objective deep reinforcement learning problem. We developed a transformer-based deep Q network to integrate individual patient disease progression and interaction effects among patients to optimize for an efficient and fair allocation policy. Our proposed model outperformed existing protocols used by different states in the U.S., by saving more lives and achieving a more equitable allocation of health resources. We refer the reader to Appendix A.7 for the limitations and future directions of this study.

Ethical Statement

The study used real patient data, which underwent Institutional Review Board (IRB) approval to ensure compliance with ethical research principles and regulatory requirements. All experiments and data processing were performed on HIPAA-compliant internal servers, ensuring the confidentiality and security of sensitive health information. The de-identified dataset used in this study will be available as a subset of a consortium-level federated dataset. Throughout the study, efforts were made to minimize bias and ensure that allocation protocols align with established ethical frameworks, promoting fairness, equity, and maximization of benefits in the allocation of health resources.

Contribution Statement

Authors Y.Li, C.M., K.H., H.W., and Z.Y. contributed equally to this work and are designated as co–first authors (marked with an asterisk * in the author list). Authors M.W. and Y.Luo, who serve as PhD advisors to the co–first authors, jointly and equally supervised this work and are recognized as co–senior authors (marked with a dagger †in the author list).

References

[Ahmad *et al.*, 2020] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3529–3530, 2020.

[Ashana *et al.*, 2021] Deepshikha Charan Ashana, George L Anesi, Vincent X Liu, Gabriel J Escobar, Christopher Chesley, Nwamaka D Eneanya, Gary E Weissman, William Dwight Miller, Michael O Harhay, and Scott D Halpern. Equitably allocating resources during crises: racial differences in mortality prediction models. *American journal of respiratory and critical care medicine*, 204(2):178–186, 2021.

[Awasthi *et al.*, 2022] Raghav Awasthi, Keerat Kaur Guliani, Saif Ahmad Khan, Aniket Vashishtha, Mehrab Singh Gill, Arshita Bhatt, Aditya Nagori, Aniket Gupta, Ponnurangam Kumaraguru, and Tavpritesh Sethi. Vacsim: Learning effective strategies for covid-19 vaccine distribution using reinforcement learning. *Intelligence-Based Medicine*, 6:100060, 2022.

[Bastani *et al.*, 2021] Hamsa Bastani, Kimon Drakopoulos, Vishal Gupta, Ioannis Vlachogiannis, Christos Hadjichristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis, and Sotirios Tsiodras. Efficient and targeted covid-19 border testing via reinforcement learning. *Nature*, 599(7883):108–113, 2021.

[Bednarski *et al.*, 2021] Bryan P Bednarski, Akash Deep Singh, and William M Jones. On collaborative reinforcement learning to optimize the redistribution of critical medical supplies throughout the covid-19 pandemic. *Journal of the American Medical Informatics Association*, 28(4):874–878, 2021.

[Bhavani *et al.*, 2021] Sivasubramanium V Bhavani, Yuan Luo, William D Miller, Lazaro N Sanchez-Pinto, Xuan Han, Chengsheng Mao, Burhaneddin Sandıkçı, Monica E Peek, Craig M Coopersmith, Kelly N Michelson, et al. Simulation of ventilator allocation in critically ill patients with covid-19. *American journal of respiratory and critical care medicine*, 204(10):1224–1227, 2021.

[Biddison *et al.*, 2019] E Lee Daugherty Biddison, Ruth Faden, Howard S Gwon, Darren P Mareiniss, Alan C Regenberg, Monica Schoch-Spana, Jack Schwartz, and Eric S Toner. Too many patients... a framework to guide statewide allocation of scarce mechanical ventilation during disasters. *Chest*, 155(4):848–854, 2019.

[Bird *et al.*, 2020] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[Chen *et al.*, 2008] Esther H Chen, Frances S Shofer, Anthony J Dean, Judd E Hollander, William G Baxt, Jennifer L Robey, Keara L Sease, and Angela M Mills. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Academic Emergency Medicine*, 15(5):414–418, 2008.

[Cimpean *et al.*, 2023] Alexandra Cimpean, Timothy Verstraeten, Lander Willem, Niel Hens, Ann Nowé, and Pieter Libin. Evaluating covid-19 vaccine allocation policies using bayesian m-top exploration. *arXiv preprint arXiv:2301.12822*, 2023.

[Dresser, 1992] Rebecca Dresser. Wanted single, white male for medical research. *The Hastings Center Report*, 22(1):24–29, 1992.

[Emanuel and Wertheimer, 2006] Ezekiel J Emanuel and Alan Wertheimer. Who should get influenza vaccine when not all can? *Science*, 312(5775):854–855, 2006.

[Emanuel *et al.*, 2020] Ezekiel J Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P Phillips. Fair allocation of scarce medical resources in the time of covid-19, 2020.

[Gostin *et al.*, 2012] Lawrence O Gostin, Kristin Viswanathan, Bruce M Altevogt, Dan Hanfling, et al. *Crisis standards of care: a systems framework for catastrophic disaster response: Volume 1: Introduction and CSC framework*, volume 3. National Academies Press, 2012.

[Grand-Clément *et al.*, 2021] Julien Grand-Clément, Carri W. Chan, Vineet Goyal, and Elizabeth Chuang. Interpretable machine learning for resource allocation with application to ventilator triage. *CoRR*, abs/2110.10994, 2021.

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[Li *et al.*, 2022] Yikuan Li, Hanyin Wang, and Yuan Luo. Improving fairness in the prediction of heart failure length of stay and mortality by integrating social determinants of health. *Circulation: Heart Failure*, 15(11):e009473, 2022.

[Liu *et al.*, 2018] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Persad *et al.*, 2009] Govind Persad, Alan Wertheimer, and Ezekiel J Emanuel. Principles for allocation of scarce medical interventions. *The lancet*, 373(9661):423–431, 2009.

[Pfohl *et al.*, 2021] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.

[Piscitello *et al.*, 2020] Gina M Piscitello, Esha M Kapania, William D Miller, Juan C Rojas, Mark Siegler, and William F Parker. Variation in ventilator allocation guidelines by us state during the coronavirus disease 2019 pandemic: a systematic review. *JAMA network open*, 3(6):e2012606–e2012606, 2020.

[Rajkomar *et al.*, 2018] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[Raschke *et al.*, 2021] Robert A Raschke, Sumit Agarwal, Pooja Rangan, C William Heise, and Steven C Curry. Discriminant accuracy of the sofa score for determining the probable mortality of patients with covid-19 pneumonia requiring mechanical ventilation. *Jama*, 325(14):1469–1470, 2021.

[Rey *et al.*, 2023] David Rey, Ahmed W Hammad, and Meead Saberi. Vaccine allocation policy optimization and budget sharing mechanism using reinforcement learning. *Omega*, 115:102783, 2023.

[Tamayo-Sarver *et al.*, 2003] Joshua H Tamayo-Sarver, Susan W Hinze, Rita K Cydulka, and David W Baker. Racial and ethnic disparities in emergency department analgesic prescription. *American journal of public health*, 93(12):2067–2073, 2003.

[Truog *et al.*, 2006] Robert D Truog, Dan W Brock, Deborah J Cook, Marion Danis, John M Luce, Gordon D Rubenfeld, Mitchell M Levy, et al. Rationing in the intensive care unit. *Critical care medicine*, 34(4):958–963, 2006.

[Van Hasselt *et al.*, 2016] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[VEN, 2015] Ventilator allocation guidelines, Nov 2015.

[Wang *et al.*, 2022] Hanyin Wang, Yikuan Li, Andrew Naidech, and Yuan Luo. Comparison between machine learning methods for mortality prediction for sepsis patients with different social determinants. *BMC Medical Informatics and Decision Making*, 22(2):1–13, 2022.

[Yu *et al.*, 2021] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[Zong and Luo, 2022] Kai Zong and Cuicui Luo. Reinforcement learning based framework for covid-19 resource allocation. *Computers & Industrial Engineering*, 167:107960, 2022.