Review

Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review

Siqi Liu[1,2], BEng; Kay Choong See[3], MBBS; Kee Yuan Ngiam[4], MBBS, MRCS, MMed, FRCS; Leo Anthony Celi[5,6], MD, MS, MPH; Xingzhi Sun[7], PhD; Mengling Feng[2], PhD

[1]NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore, Singapore
[2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore
[3]Division of Respiratory & Critical Care Medicine, National University Hospital, Singapore, Singapore
[4]Group Chief Technology Office, National University Health System, Singapore, Singapore
[5]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States
[6]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States
[7]Ping An Health Technology, Beijing, China

Corresponding Author:
Mengling Feng, PhD
Saw Swee Hock School of Public Health
National University of Singapore
12 Science Drive 2, #10-01
Singapore, 117549
Singapore
Phone: 65 65164988
Email: ephfm@nus.edu.sg

Abstract

Background: Decision support systems based on reinforcement learning (RL) have been implemented to facilitate the delivery of personalized care. This paper aimed to provide a comprehensive review of RL applications in the critical care setting.

Objective: This review aimed to survey the literature on RL applications for clinical decision support in critical care and to provide insight into the challenges of applying various RL models.

Methods: We performed an extensive search of the following databases: PubMed, Google Scholar, Institute of Electrical and Electronics Engineers (IEEE), ScienceDirect, Web of Science, Medical Literature Analysis and Retrieval System Online (MEDLINE), and Excerpta Medica Database (EMBASE). Studies published over the past 10 years (2010-2019) that have applied RL for critical care were included.

Results: We included 21 papers and found that RL has been used to optimize the choice of medications, drug dosing, and timing of interventions and to target personalized laboratory values. We further compared and contrasted the design of the RL models and the evaluation metrics for each application.

Conclusions: RL has great potential for enhancing decision making in critical care. Challenges regarding RL system design, evaluation metrics, and model choice exist. More importantly, further work is required to validate RL in authentic clinical environments.

(J Med Internet Res 2020;22(7):e18477) doi: 10.2196/18477

KEYWORDS

artificial intelligence; reinforcement learning; critical care; decision support systems, clinical; intensive care unit; machine learning

Introduction

Background

In the health care domain, clinical processes are dynamic because of the high prevalence of complex diseases and dynamic changes in the clinical conditions of patients. Existing treatment recommendation systems are mainly implemented using rule-based protocols defined by physicians based on evidence-based clinical guidelines or best practices [1-3]. In addition, these protocols and guidelines may not consider

multiple comorbid conditions [4]. In an intensive care unit (ICU), critically ill patients may benefit from deviation from established treatment protocols and from personalizing patient care using means not based on rules [5,6].

When physicians need to adapt treatment for individual patients, they may take reference from randomized controlled trials (RCTs), systemic reviews, and meta-analyses. However, RCTs may not be available or definitive for many ICU conditions. Many patients admitted to ICUs might also be too ill for inclusion in clinical trials [6]. Furthermore, only 9% of treatment recommendations in the ICU are based on RCTs [7], and the vast majority of RCTs in critical care have negative findings [8]. To aid clinical decisions in ICUs, we need other methods, including the use of large observational data sets. ICU data can be useful for learning about patients as they were collected in a data-rich environment. A large amount of data can then be fed into artificial intelligence (AI) systems (using computers to mimic human cognitive functions) and machine learning methods (using computer algorithms to perform clinical tasks without the need for explicit instructions). AI and machine learning can then help with diagnosis [9,10], treatment [11,12], and resource management [13,14] in the ICU. Given the dynamic nature of critically ill patients, one machine learning method called reinforcement learning (RL) is particularly suitable for ICU settings.

Fundamentals of Reinforcement Learning

RL is a goal-oriented learning tool where a computer *agent*, acting as a decision maker, analyzes available data within its defined environment [15], derives a rule for taking actions, and optimizes long-term rewards. The agent is the RL model that we wish to develop. In general, an RL agent receives evaluative feedback about the performance of its action in each time step, allowing it to improve the performance of subsequent actions by trial and error [16]. Mathematically, this sequential decision-making process is called the Markov decision process (MDP) [17]. An MDP is defined by 4 major components: (1) a state that represents the environment at each time; (2) an action the agent takes at each time that influences the next state; (3) a transition probability that provides an estimate for reaching different subsequent states, which reflects the environment for an agent to interact with; (4) a reward function is the observed feedback given a state-action pair. The solution of the MDP is an optimized set of rules and is termed the policy.

RL has already emerged as an effective tool to solve complicated control problems with large-scale, high-dimensional data in some application domains, including video games, board games, and autonomous control [18-20]. In these domains, RL has been proven to achieve human-level capacity for learning complex sequential decisions. For instance, Alpha Go is an RL agent for playing the strategy board game Go. On the basis of Alpha Go's learned policy, and given the current position of the Go stones, it is possible to decide where the next white/black stone should be placed on the board to maximize its chance of winning.

Analogies to Critical Care

For critical care, given the large amount and granular nature of recorded data, RL is well suited for providing sequential treatment suggestions, optimizing treatments, and improving outcomes for new ICU patients. RL also has the potential to expand our understanding of existing clinical protocols by automatically exploring various treatment options. The RL agent analyzes the patient trajectories, and through trial and error, derives a policy, a personalized treatment protocol that optimizes the probability of favorable clinical outcomes (eg, survival). As this computerized process is an attempt to mimic the human clinician's thought process, RL has also been called the AI clinician [21].

We can consider the state as the well-being/condition of a patient. The state of the patients could depend on static traits (eg, patient demographics including age, gender, ethnicity, pre-existing comorbidity) and longitudinal measurements (eg, vital signs, laboratory test results). An action is a treatment or an intervention that physicians do for patients (eg, prescription of medications and ordering of laboratory tests). The transition probability is the likelihood of state transitions, and it is viewed as a prognosis. If the well-being in the new state is improved, we assign a reward to the RL agent, but we penalize the agent if the patient's condition worsens or stays stagnant after the intervention.

As illustrated in Figure 1, if we take a snapshot of the current well-being of a patient as his/her state, the physician would provide a treatment or an intervention (an action) to the patient. This action would lead the patient to the next state depending on his/her current state and the action performed on him/her. While knowing the next state of the patient, the physician would need to take another action according to the new state. These state-action pairs would continue to rollout over time, and the resultant trajectory of state-action pairs could represent the changes in the patients' conditions and the sequential treatment decisions that were performed by the physicians. We can define the length of the trajectory for each patient as fixed (eg, during the first 24 hours of the ICUs stay) or as dynamic (eg, different patients could be discharged from the ICUs at different times).

Figure 1. Illustration of reinforcement learning in critical care.

The main objective of the RL algorithm is to train an agent that can maximize the cumulative future reward from the state-action pairs given the patients' state-action trajectories. When a new state is observed, the agent is able to perform an action, which could choose the action for the greatest long-term outcome (eg, survival). When the RL agent is well-trained, it is possible to pick the best action given the state of a patient, and we describe this process as acting according to an optimal policy.

A policy is analogous to a clinical protocol. Nonetheless, a policy has advantages over a clinical protocol because it is capable of capturing more personalized details of individual patients. A policy can be represented by a table where it maps all possible states with actions. Alternatively, a policy could also be represented by a deep neural network (DNN) where given the input of a patient's state, the DNN model outputs the highest probability of an action. An optimal policy can be trained using various RL algorithms. Some widely applied RL algorithms include the fitted-Q-iteration (FQI) [22], deep Q network (DQN) [23], actor-critic network [24], and model-based RL [25]. More technical details about various RL models have been explained [26,27].

As RL in critical care is a relatively nascent field, we therefore aimed to review all the existing clinical applications that applied RL in the ICU setting for decision support over the past 10 years (2010-2019). Specifically, we aimed to categorize RL applications and summarize and compare different RL designs. We hope that our overview of RL applications in critical care can help reveal both the advances and gaps for future clinical development of RL. A detailed explanation of the concept of RL and its algorithms is available in Multimedia Appendix 1 [28].

Methods

Search Strategy

A review of the literature was conducted using the following 7 databases: PubMed, Institute of Electrical and Electronics Engineers (IEEE), Google Scholar, Medical Literature Analysis and Retrieval System Online (MEDLINE), Excerpta Medica Database (EMBASE), ScienceDirect, and Web of Science. The search terms *reinforcement learning*, *critical care*, *intensive care*, *intensive care units*, and *ICUs* were combined. The search phrases listed in Textbox 1 were used to identify articles in each database.

Textbox 1. Queries used to retrieve records.

EMBASE (Excerpta Medica Database)

- #1 'reinforcement learning'
- #2 'intensive care unit' OR 'critical care' OR 'ICU'
- #1 AND #2

Google Scholar

- (conference OR journal) AND ("intensive care unit" OR "critical care" OR ICU) AND "reinforcement learning" -survey -reviews -reviewed -news

IEEE (Institute of Electrical and Electronics Engineers)

- (("Full Text Only": "reinforcement learning") AND "Full Text Only": "intensive care units") OR (("Full Text Only": "reinforcement learning") AND "Full Text Only": "critical care")

MEDLINE (Medical Literature Analysis and Retrieval System Online)

- multifield search=reinforcement learning, critical care, intensive care

PubMed

- ("reinforcement learning") AND (("ICU") OR ("critical care") OR ("intensive care unit") OR ("intensive care"))

ScienceDirect

- "reinforcement learning" AND ("critical care" OR "intensive care" OR "ICU")

Web of Science

- ALL=(intensive care unit OR "critical care" OR "ICU") AND ((ALL=("reinforcement learning")) AND LANGUAGE: (English))

Inclusion Criteria

To be eligible for inclusion in this review, the primary requirement was that the article needed to focus on the implementation, evaluation, or use of an RL algorithm to process or analyze patient information (including simulated data) in an ICU setting. Papers published from January 1, 2010, to October 19, 2019 were selected. General review articles and articles not published in English were excluded. Only papers that discussed sufficient details on the data, method, and results were included in this review.

Data Synthesis

Data were manually extracted from the articles included in the review. A formal quality assessment was not conducted, as relevant reporting standards have not been established for articles on RL. Instead, we extracted the following characteristics from each study: the purpose of the study, data source, number of patients included, main method, evaluation metrics, and related outcomes. The final collection of articles was divided into categories to assist reading according to their application type in the ICUs.

Results

Selection Process and Results Overview

The selection process of this review was demonstrated using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram (Figure 2). From the full text of 269 distinct articles, an independent assessment for eligibility was performed by 2 authors (SL and MF). Disagreements were discussed to reach consensus. During the full-text review, 249 articles were excluded, and 21 articles were eventually included. The reasons for exclusion during the review process are outlined in Table 1.

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram of the search strategy.

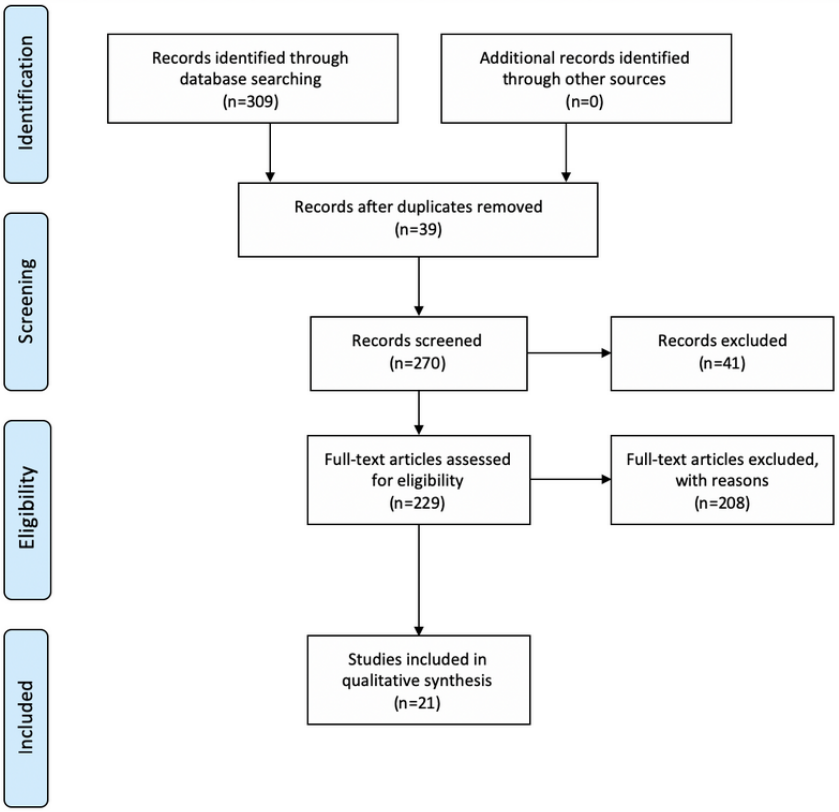


Table 1. Exclusion criteria used to exclude papers.

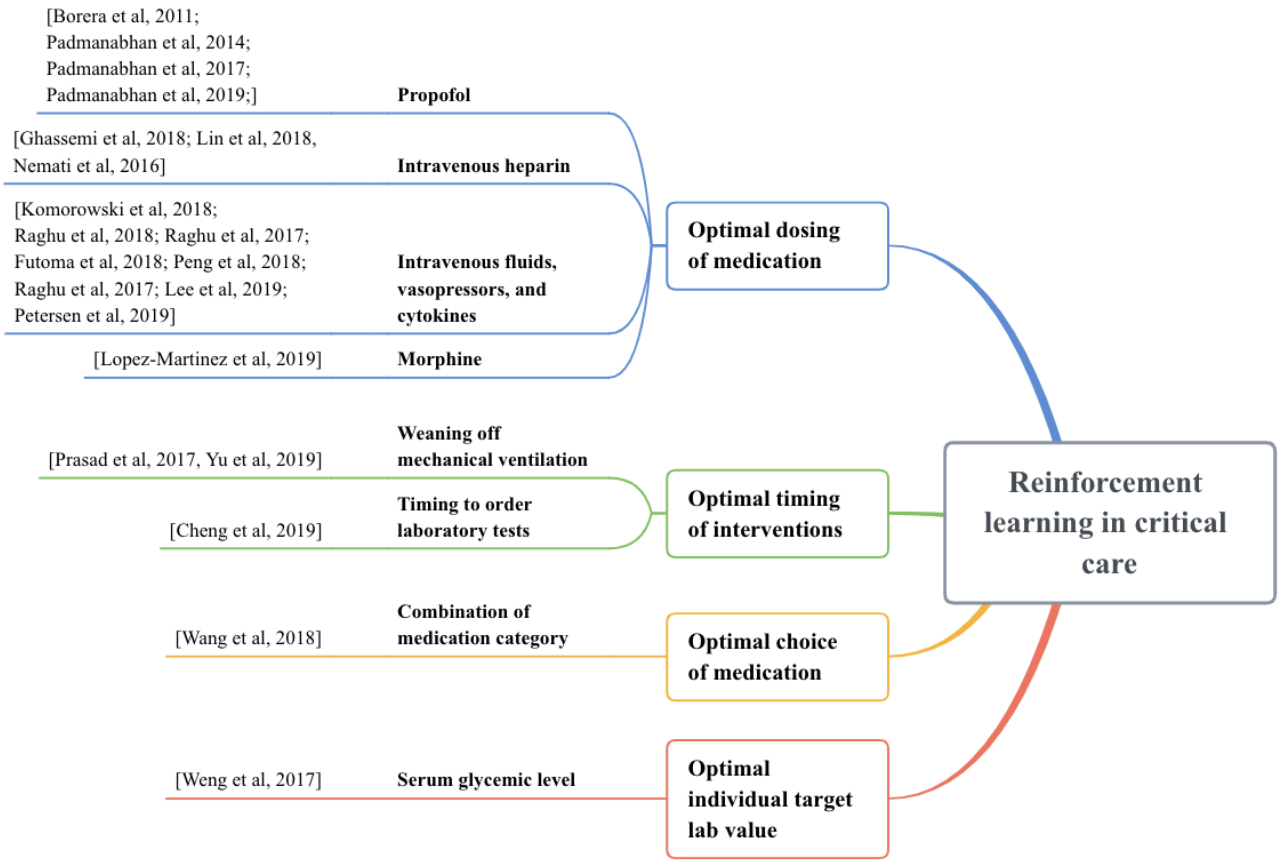
Criterion number	Exclusion criteria	Justification	Excluded articles, n
1	Duplicates	The papers have duplicate titles	39
2	Not a research article	The papers were blog articles, reports, comments, or views	23
3	Not written in English	The papers were not written in English	6
4	Review	The papers were review articles regarding general methods on big data, deep learning, and clinical applications	12
5	Not applied in the field of critical care	The papers did not focus on applications in critical care or intensive care	92
6	Not using RL[a] as the approach in critical care	The papers discussed issues in the critical care setting, but not using RL as an approach	115
7	No clear description of the method and result	The methods and results were not clearly described and thus not qualified for this review	1

[a]RL: reinforcement learning.

In this section, we organized the reviewed articles into 4 categories, which reflect clinically relevant domains: (1) optimal individualized target laboratory value; (2) optimal choice of medication; (3) optimal timing of an intervention; and (4) optimal dosing of medication.

We plotted the number of articles reviewed by their category and year of publication in Figure 3. We found that the majority of the papers were published in the past 3 years (n=17), indicating an increasing trend of applying RL-based approaches to assist physicians in decision making in critical care. In each of the 4 categories, we further organized the articles into subgroups based on their clinical questions (Figure 3). The figure shows that most of the applications used RL to find optimal drug dosing (n=16) [6,21,29-42], followed by the timing of an intervention (n=3) [43-45]. Only a few applications were looking at the individualized laboratory value (n=1) [46] and the optimal choice of medication (n=1) [47].

Figure 3. Mapping of reinforcement learning studies in critical care by application type.



Next, we discuss the details for each category with the methods and outcomes for each application. In particular, we further grouped the studies based on specific medication or treatment type in categories 3 and 4 to assist readers. A summary of all study details is found in Multimedia Appendix 2.

Optimal Individualized Target Laboratory Value

Even after decades of routine use of laboratory value ranges, reference standards may need to be reconsidered, especially for individual patients [48]. Personalized targets for laboratory values in ICU patients could account for disease severity, comorbidities, and other patient-specific differences. Weng et al [46] tried to identify individualized targeted blood glucose levels as a reference for physicians. They applied an RL-based approach, *policy iteration*, to learn the target glycemic range at an hourly interval for severely ill patients with sepsis using real ICU data. Their approach was tested using the Medical Information Mart for Intensive Care III (MIMIC III), a large, publicly available ICU database [49]. MIMIC III contains information for hospital admissions of 43,000 patients in critical care units during 2001 and 2012, from which the authors extracted hourly data for 5565 patients with sepsis.

Weng et al [46] constructed their RL model as follows: First, they represented the patients' states from 128 variables. These variables included patient demographics, comorbid conditions, vital sign changes, and laboratory value changes. They used a spare autoencoder [50] to reduce the high dimensionality of the raw features (128 dimensions) to only 32 dimensions so that the RL model could be trained more efficiently with limited observational data. Second, they chose to act upon 1 of 11 discrete ranges of serum glucose at each time step. Third, they designed the reward function so that the RL agent could recommend an hourly target glucose level to optimize long-term survival. A positive 100 was assigned to the end state if patients survived 90 days after admission, and a negative 100 was assigned if the patients died. For each state-action pair, the *value* of the pair was iteratively estimated using the reward from the training data.

To understand how the reward value was related to mortality, the authors assigned values to discrete buckets using separate test data. In each value bucket, if the state-action pair is part of a trajectory where a patient died, a label of 1 was assigned to that bucket; otherwise, a label of 0 was assigned. After assigning all the state-action pairs from the test data with the labels in the corresponding value bucket, the mortality rate could be estimated for each value bucket. The authors plotted the estimated mortality rate with respect to the value-buckets and found an inverse relationship between them, where the highest value was associated with the lowest mortality. This result suggested that the learnt value represented the relationship between the state-action pair and mortality and that the learnt value of the state-action pairs from training data was validated on the test data.

To validate the RL policy, the author calculated the frequency of state transitions from the training data and generated new

trajectories. Starting from the observed state in the test data, the RL policy would recommend an action with the highest value, and the subsequent state was estimated with the transition probability. By averaging the value for all state-action pairs in the simulated trajectory, the mortality for simulated trajectories could be estimated by mapping this value in the mortality-value plot. Compared with the actual mortality rate in the test data, the author claimed that if physicians could control patients' hourly blood glucose levels within the range recommended by the RL model, the estimated 90-day mortality would be lowered by 6.3% (from 31% to 24.7%).

Optimal Choice of Medications

Apart from some clinical decision support systems, commonly used systems such as computerized prescriber order entry and bar-coded medication administration lack personalized recommendations to optimize medication effectiveness and minimize side effects [51]. Wang et al [47] applied a deep learning network based on RL to exploit medication recommendations with a data-driven strategy. Their approach accounted for individual patient demographics, laboratory values, vital signs, and diagnoses from the MIMIC III database. They selected the top 1000 out of 4127 medications and the top 2000 out of 6695 diseases (represented by the International Classification of Diseases, Ninth Revision codes), which covered 85.4% of all medication records and 95.3% of all diagnosis records, respectively. To reduce the problem complexity, the authors further categorized the 1000 medications into 180 drug categories using anatomical therapeutic chemical codes and aggregated patients' drug prescriptions into 24-hour windows.

The authors defined RL action as the medication combinations from the 180 drug categories. They adopted an actor-critic RL agent that suggested a daily medication prescription set, and aimed to improve patients' hospital survival. The details of the actor-critic RL algorithm are explained in Multimedia Appendix 1 [28]. For each patient's ICU day, the *actor* network would recommend one medication combination by considering state variables such as demographics, laboratory results, and vital signs. A reward value of positive 15 would be given to the end state if a patient survived until hospital discharge and negative 15 if the patient died. The reward was designated as 0 for all other time steps. To counterbalance the *actor* network, the *critic* network was applied to evaluate the consistency of actual physician prescriptions and the RL agent's recommendations. The net effect of the actor-critic RL agent was to optimize the long-term outcomes of patients (hospital mortality) while minimizing deviations of RL-recommended actions from actual prescription patterns. In addition to the actor-critic network, the authors also applied long short-term memory [52] to represent a patient's current state by incorporating the long sequence of all historical states. Wang et al [47] suggested that hospital mortality would be reduced by 4.4% if clinicians adhered to the RL agent's recommendations.

Optimal Timing of Intervention

Weaning of Mechanical Ventilation

Mechanical ventilation (MV) is a life-saving treatment applied in approximately a third of all critically ill patients [53]. Prematurely discontinuing MV (premature weaning) and excessively prolonged MV (late weaning) are both associated with higher mortality [54]. The best time to wean may be uncertain [55].

To optimize the timing of ventilation discontinuation, Prasad et al [43] applied the RL-based FQI (the details of the FQI algorithm are explained in Multimedia Appendix 1 [28]) on the MIMIC III database for all patients who were kept under ventilator support for more than 24 hours and extracted their records every 10 min from ICU admission to discharge. Patient states included a number of factors that could affect extubation, such as demographics, pre-existing conditions, comorbidities, and time-varying vital signs. The action for the ventilation setting was binary, that is, for each 10-min time step, the RL agent needed to decide whether the ventilation should be set on (continued MV) or off (weaned from MV). For reward design, Prasad et al [43] followed an existing weaning protocol from the Hospital of University of Pennsylvania. They assigned reward values to the RL agent at each time step according to 3 major considerations: (1) the RL agent should penalize each additional hour spent on the ventilator, (2) the RL agent should be assigned a positive reward value to a weaning action if the patient's vital signs and laboratory results were steady and within normal ranges after extubation, and (3) there was no reward value for failed spontaneous breathing trial or for reintubation after the first extubation. For RL policy evaluation, the authors calculated the proportion of weaning actions from the RL policy, referencing the total number of weaning actions from the clinician's policy at each time step, and calculated the overall consistency of weaning transitions. The recommend actions from the RL agent could match 85% of those from clinicians. The authors categorized the degree of consistency into 5 bins, and plotted the distribution of the number of reintubations with respect to the discrete consistency levels. Their results showed that when the consistency was high, vital sign fluctuations were fewer, laboratory results were more in-range, and reintubations were minimized.

Yu et al [45] studied the same clinical issue as Prasad et al [43] and used the same data set, but designed a different reward function using inverse RL. The inverse RL model directly learnt reward mapping from data for each state-action pair and inferred what clinicians would wish to achieve as a reward. Similar to Prasad et al [43], the RL recommendations by Yu et al [45] were associated with shorter weaning times and fewer reintubations compared with clinician decision making.

Timing to Order Laboratory Tests

The timing of ordering a laboratory test can be challenging. Delayed testing would lead to continued uncertainty over the patient's condition and possible late treatment [56]. However, excessively early ordering of laboratory tests can cause unnecessary discomfort to the patient, increase the risk of anemia, and increase health care cost.

Cheng et al [44] applied the FQI method to find the optimal timing for ordering laboratory tests among patients with sepsis in the MIMIC III data set. They examined the timing of 4 types of laboratory tests: white blood cell count (WBC), creatinine, blood urea nitrogen (BUN), and lactate. They sampled the

patients' data at hourly intervals and constructed the state of a patient by considering the predictive variables of severe sepsis or acute kidney failure, including respiratory rate, heart rate, mean blood pressure, temperature, creatinine, BUN, WBC, and lactate. The missing values were predicted by a multioutput Gaussian process [57,58]. In their RL model, they chose to design the reward function with the combination of 4 factors: (1) a positive reward should be given only if the ordering of test was necessary, while penalizing over or under ordering; (2) the RL agent should be encouraged to order laboratory tests when there was a sudden change in laboratory results or vital signs; (3) negative reward should be given if the laboratory results were similar to the last measurements (no information gain); (4) a penalty would be added to a reward whenever a test was ordered, to reflect the testing cost. Their RL agent, compared with clinicians, was able to reduce the number of laboratory tests by 27% for lactate and 44% for WBC, while maintaining high information gain.

Optimal Dosing of a Drug

Recommendations for dosing regimens in ICU patients are often extrapolated from clinical trials in healthy volunteers or noncritically ill patients. This extrapolation assumes similar drug behavior (pharmacokinetics and pharmacodynamics) in the ICU and other patients or healthy volunteers. However, it is well known that many drugs used in critically ill patients may have alterations in pharmacokinetic and pharmacodynamic properties because of pathophysiological changes or drug interactions [59]. Therefore, critically ill patients bring unique challenges in drug dosing.

Dosing of Propofol

Critically ill patients in ICUs often require sedation to facilitate various clinical procedures and to comfort patients during treatment. Propofol is a widely used sedative medication [60], but titration of propofol is challenging, and both over sedation and under sedation can have adverse effects [32]. Of the studies reviewed, 6 studies have focused on applying RL to determine the optimal dosage for propofol while maintaining the physiological stability of the patient. The bispectral index (BIS) was used to monitor sedation level and to determine the effect of propofol.

Borera et al [29] was the first to apply RL to a pharmacokinetic model [61] to describe the time-dependent distribution of propofol in human surgical patients. The RL agent was a neural network aimed at optimizing the propofol dose to achieve the target BIS value. The patient's state and state transition were modeled using a mathematical pharmacokinetic model with predefined parameters such as the concentration at half maximal effect of BIS, degree of nonlinearity of BIS, and time-lag coefficient to estimate the BIS value for simulated patients. The action was a discrete range of propofol infusion rate. The reward function was the error rate between the target BIS value and the current simulated BIS value, where a larger negative reward was given when the current simulated BIS value was further away from the predefined target value. They measured the performance of the RL agent by looking at the time to reach the target BIS value (steady time). The evaluation was conducted on 1000 simulated patients. On average, the steady time was 3.25 min for the BIS value to reach target.

To ensure patient safety, propofol dosing should consider the concurrent stability of vital parameters. For instance, Padmanabhan et al [30] chose mean arterial pressure (MAP) as the secondary control variable. The authors combined the error rates for both BIS and MAP when designing the reward. The target for the RL agent was to infuse propofol so that the target BIS would be reached in a short time, whereas MAP was kept within a desired range. In subsequent studies, Padmanabhan et al [31,32] modified their methods with different RL training algorithms (Q-learning and policy iteration). In all their studies, the RL agent was able to suggest accurate propofol doses and achieve target BIS values within a few minutes.

In contrast to fixed pharmacokinetic models in the RL model environment, Yu et al [45] applied FQI and Bayesian inverse RL on the MIMIC III database. They considered patients' demographic characteristics, pre-existing conditions, comorbidities, and time-varying vital signs to construct the state of the patient. Their inverse RL model interpreted clinician preference as a reward for different patient states. The learned reward function from the inverse RL model suggested that clinicians may pay more attention to patients' cardiorespiratory stability rather than oxygenation when making decisions about propofol dosage.

Dosing of Intravenous Heparin

Anticoagulant agents are often used to prevent and treat a wide range of cardiovascular diseases. Heparin is commonly used in critical care [62], yet its precise dosing is complicated by a narrow therapeutic window. Overdosing of heparin results in bleeding whereas under dosing risks clotting. To guide heparin dosing, activated partial thromboplastin time (aPTT) is often used as a measure of the anticoagulant effect of heparin.

Nemati et al [6] applied FQI with a neural network to optimize and individualize heparin dosing. Their study was conducted on the MIMIC II database, with the reward function based on aPTT levels following heparin dosing [63]. The reward to the RL agent will be high if the aPTT value is between 60 and 100 seconds. After training, they plot the state-action value with respect to the level of consistency between the RL policy and clinician practice. Their results showed that, on average, following the recommendations of the RL agent resulted in higher state-action values.

Ghassemi et al [33] and Lin et al [34] focused on a personalized optimal heparin dosing using different RL algorithms. In addition to the MIMIC III data set, Lin et al [34] applied an actor-critic network on the Emory Healthcare data set from Emory University. For RL policy evaluation, Lin et al [34] regressed the discordance between RL policy and physician practice over the number of clotting and bleeding complications, adjusting for covariates such as history of clot or bleed, weight, age, and sequential organ failure assessment score. The regression coefficient suggested that following the RL agent's recommendations would have likely resulted in improved clinical outcomes with a reduced number of clotting and bleeding complications.

Intravenous Fluids, Vasopressors, and Cytokine Therapy for Treating Sepsis

Sepsis is the third leading cause of death and is expensive to treat [64]. Besides antibiotics and source control, challenges remain with the use of intravenous (IV) fluids to correct hypovolemia and administration of vasopressors to counteract sepsis-induced vasodilation. Raghu et al [36] suggested a data-driven RL approach to recommend personalized optimal dosage for IV fluids and vasopressors to improve hospital mortality. Their RL model was double DQN with dueling, which can minimize the overestimation problem of previous Q-learning models. The details of the Q-learning and double DQN algorithms are explained in Multimedia Appendix 1 [28]. The authors considered patients' demographics, laboratory values, vital signs, and intake/output events as state features in the RL model. Action was designed as a combination of 5 discrete bins for IV fluid dosing and 5 bins for vasopressor dosing to treat patients with sepsis. The reward was issued at the terminal time step of the patient's trajectory, with a positive reward if the patient survived. Data were extracted from the MIMIC III database for all patients who fulfilled sepsis-3 criteria [65]. For policy evaluation, Raghu et al [36] plotted the estimated hospital mortality with respect to the difference between dosages recommended by the RL agent and by clinicians. The plot showed that the mortality was lowest when there was no discrepancy between RL policy and physician decision making. Six other groups of researchers also focused on the same research question and applied various RL algorithms with slightly different designs of the state space, reward function, and evaluation metrics [21,35,37-40]. The findings from these studies all suggest that the RL agent would be able to learn from the data and if physicians followed the RL policy, the estimated hospital mortality could be improved.

Among the aforementioned studies, Komorowski et al [21] were the pioneers of applying RL in the ICU, using data from patients with sepsis in the MIMIC III database. They inferred a patient's health status using an array of inputs, which included demographics, vital signs, laboratory tests, illness severity scores, medications, procedures, fluid intake and output, physician notes, and diagnostic coding. Patient data were aggregated and averaged every 4 hours to represent patient states. Using a k-means algorithm, these patient states were then simplified into 750 discrete mutually exclusive clusters. A sequence of these clustered states would describe a particular patient's trajectory. The authors estimated the state transition probability by counting how many times each transition was observed and converted the counts to a stochastic matrix. This transition matrix contained the probability for each patient going to a new state, given a previous action taken in the current state. The entire trajectory of a patient's state can be estimated using the transition matrix. The authors applied a policy iteration RL algorithm that learnt the optimal dosing policy for IV fluids and vasopressors to maximize the probability of 90-day survival.

Nevertheless, the study by Komorowski et al [21]. had several limitations. First, their study only considered fluid and vasopressor management, ignoring other important treatments such as source control, correction of hypovolemia, and management of secondary organ failures [21]. Second, 90-day mortality is affected by factors outside of the ICU, which the study did not take into account. Third, clinical decision making considers both short-term outcomes (eg, physiological stability) and long-term outcomes (eg, kidney failure or mortality), but the study only considered mortality as the single goal for training the RL algorithm [66]. Fourth, discretizing patient health status into discrete clusters loses data granularity and may limit the ability to detect changes in patient status. These limitations also occur in other studies, which we will elaborate in the Discussion section.

Other than using IV fluids and vasopressors for treating sepsis. Petersen et al [42] investigated cytokine therapy using the deep deterministic policy gradient [67] method. The details of the policy gradient RL algorithm are explained in Multimedia Appendix 1 [28]. They evaluated the RL model by using an agent-based model, the innate immune response agent-based model [68], that simulated the immune response to infection. The RL policy was able to achieve a very low mortality rate of 0.8% over 500 simulated patients, and suggested that personalized multicytokine treatment could be promising for patients with sepsis.

Dosing of Morphine

Critically ill patients may experience pain as a result of disease or certain invasive interventions. Morphine is one of the most commonly used opioids for analgesia [69]. Similar to sedation, the dosing of analgesia is subject to uncertainty. Lopez-Martinez et al [41] collected data for patients who had at least one pain intensity score and at least one dose of IV morphine in the MIMIC III database. They applied double DQN with dueling as their RL model and constructed the state space to be continuous with features including the patient's self-reported pain intensity and their measured physiological status. The action was a choice of 14 discrete dosing ranges of IV morphine. The reward was determined by considering both the patients' cardiorespiratory stability and their pain intensity. The highest reward was given when pain was absent and both heart rate and respiration rate were within the acceptable range. By comparing the RL policy with physicians' choices, Lopez-Martinez et al [41] found that RL policy tended to prescribe higher doses of morphine. This result was consistent with previous studies: continuous dosing provided similar or even better pain relief with no increase in acute adverse effects [70,71].

Discussion

Principal Findings

Our comprehensive review of the literature demonstrates that RL has the potential to be a clinical decision support tool in the ICU. As the RL algorithm is well aligned with sequential decision making in ICUs, RL consistently outperformed physicians in simulated studies. Nonetheless, challenges regarding RL system design, evaluation metrics, and model choice exist. In addition, all current applications have focused on using retrospective data sets to derive treatment algorithms and require prospective validation in authentic clinical settings.

RL System Design

The majority of applications were similar in their formulation of the RL system design. The state space is usually constructed by features including patient demographics, laboratory test values, and vital signs, whereas some studies applied encoding methods to represent the state of the patients instead of using raw features. The action space was very specific to each application. For instance, in terms of the dosing category, the action space would be discretized ranges of medication dosage. For other categories, such as timing of an intervention, the action space would be the binary indicator of an intervention for each time step. The number of action levels differed among the studies. For some studies, the action levels could be as many as a dozen or a hundred (eg, optimal medication combination), whereas for other studies, the action levels were limited to only 2 (eg, on/off MV). The design of the reward function is central to successful RL learning. Most of the reward functions were designed a priori with guidance from clinical practice and protocols, but 2 studies [40,45] managed to directly learn the reward function from the data using inverse RL.

Evaluation Metrics

The only metric that matters is if the adoption of an RL algorithm leads to improvement in some clinical outcomes. Most studies calculated the estimated mortality as the long-term outcome and drew plots to show the relationship between the estimated mortality versus the learnt value of patients' state-action trajectories, where the higher value function was associated with lower mortality. The RL agent would provide treatment suggestions for those actions with higher values, thus leading to a lower estimated mortality. Estimated mortality is a popular metric for RL policy evaluation. However, the problem with the estimated mortality is that it is calculated from simulated trajectories with observational data, and may not be the actual mortality.

Mortality is not always the most relevant and appropriate outcome measure. For instance, in the study by Weng et al [46], they tried to identify individualized targeted blood glucose levels as a reference for physicians. In their study, 90-day mortality was used to evaluate the RL policy. However, a more relevant measure could be considered, such as short-term changes in the blood glucose level, physiological stability, and development of complications.

Several studies that focused on propofol titration have considered BIS as the evaluation metric to monitor the sedation level and hence to determine the effect of propofol. Although BIS monitoring is fairly objective, assessing sedation is usually performed by health care providers with clinically validated behavioral assessment scales such as the Richmond Agitation-Sedation Scale score [72]. In addition, EEG-based technologies, such as BIS and M-entropy, have been validated more in the operating room than in the ICU [73]. Furthermore, BIS cannot be used as the sole monitoring parameter for sedation, as it is affected by several other factors, including the anesthetic drugs used, muscle movement, or artifacts from surgical equipments [74].

To date, there has been no prospective evaluation of an RL algorithm. Moreover, the observational data itself may not truly reflect the underlying condition of patients. This is known as the partially observable MDP [75] problem, where we are only able to represent a patient's state by the observed physiological features, which are solved by mathematical approximation.

Model Choice

FQI and DQN seem to be the top RL approaches among the reviewed studies. FQI is not a deep learning–based RL model, which guarantees convergence for many commonly used regressors, including kernel-based methods and decision trees. On the other hand, DQN leverages the representational power of DNNs to learn optimal treatment recommendations, mapping the patient state-action pair to the value function. Neural networks hold an advantage over tree-based methods in iterative settings in that it is possible to simply update the network weights at each iteration, rather than rebuilding the trees entirely.

Both FQI and DQN are off-policy RL models. Off-policy refers to learning about one way of behaving from the data generated by another way of selecting actions [76]. For instance, an off-policy RL model tries to train a policy X to select actions in each step, but it estimates the Q-values from state-action pairs where the action was chosen by following another policy Y. In contrast to off-policy learning, on-policy learning uses the same policy X to choose actions and to evaluate the returns in each step during training. Most of the included studies adopted off-policy RL models because the RL models aim to learn policy X from the data, which was generated by following real actions of physicians (policy Y). The data generated by policy Y is the actual physicians' policy, where the RL models try to learn and improve from. This is the fundamental idea of applying off-policy RL models.

In addition, both FQI and DQN are value-based RL models that aim to learn the value functions. In value-based RL, a policy can be derived by following the action with the highest value at each time step. Another type of RL is called policy-based RL, which aims to learn the policy directly without worrying about the value function. Policy-based methods are more useful in continuous space. When the data volume is insufficient to train a DQN model, the DQN is not guaranteed to achieve a stable RL policy. As there is an infinite number of actions or states to estimate the values for, value-based RL models are too computationally expensive in the continuous space. However, policy-based RL models demand more data samples for training. Otherwise, the learned policy is not guaranteed to converge to an optimal one. Both value-based and policy-based RL models can be grouped in a more general way as *model-free* RL. Here the word *model-free* means the environment is unknown to an agent. The RL agent makes use of the trajectories generated from the environment, rather than explicitly knowing the rule or the transition probability. In contrast to model-free RL, *model-based* RL requires the agent to know the transition probability for all the state-action combinations explicitly and hence impractical as the state space and action space grow. In the critical care context, patients' conditions and prognosis are very complex to apply model-based RL because we are not exactly sure about the probability of all state transitions. In

addition, most studies in critical care could only use limited retrospective data to train the model offline. Therefore, we found that most of the studies have applied a value-based RL model to utilize the available observational data.

Common Data Sets

We found that 71% (15/21) of applications utilized the MIMIC II or MIMIC III database to conduct their experiments. We conjecture that such popularity might be due to public availability and high quality of MIMIC data. However, data collected from a single source may introduce potential bias to the research findings. There are inherent biases in the medical data sets obtained at various institutions due to multiple factors, including operation strategy, hospital protocol, instrument difference, and patient preference. Therefore, the RL models trained on a single data set, regardless of the data volume, cannot be confidently applied to another data set. The findings from the reviewed articles may not be generalizable to other institutions and populations. In addition to the MIMIC database, one of the studies also utilized the eICU Research Institute (eRI) database to test their RL model [77]. The eRI database has a larger volume of data compared with the MIMIC database, and it is also publicly available. We suggest that future applications could cross-validate their models on both the MIMIC and eRI databases. In addition, all current applications have focused on using retrospective data sets to derive treatment algorithms and require prospective validation in authentic clinical settings.

Strengths and Limitations of This Study

The strengths of this paper include the comprehensive and extensive search for all available publications that applied RL as an approach in the critical care context. Nonetheless, we acknowledge the limitations. We included papers (eg, those on arXiv) that have not been peer-reviewed *before* publication but these papers have undergone a postpublication peer review. According to the search phrases applied in this review, we may miss out certain papers that applied RL in critical care, but did not specify the phrase *intensive care* nor *ICU* in their full text papers.

Challenges and Future Directions

A number of challenges must be overcome before RL can be implemented in a clinical setting. First, it is important to have a meaningful reward design. The RL agent would be vulnerable in case of reward misspecification, and might not be able to produce any meaningful treatment suggestion. Inverse RL can be an alternative to a priori–specified reward functions. However, inverse RL assumes that the given data represent the experts' demonstrations and the recommendations from the data were already optimal; these may not be true.

Second, medical domains present special challenges with respect to data acquisition, analysis, interpretation, and presentation of these data in a clinically relevant and usable format. Addressing the question of censoring in suboptimal historical data and explicitly correcting for the bias that arises from the timing of interventions or dosing of medication is crucial to fair evaluation of learnt policies.

Third, another challenge for applying the RL model in the clinical setting is exploration. Unlike other domains such as game playing, where one can repeat the experiments as many times, in the clinical setting, the RL agent has to learn from a limited set of data and intervention variations that were collected offline. Using trial and error to explore all possible scenarios may conflict with medical ethics, thereby limiting the ability of the RL agent to attempt new behaviors to discover ones with higher rewards and better long-term outcomes.

In comparison with other machine learning approaches, there is an absence of acceptable performance standards in RL. This problem is not unique to RL but seems harder to address in RL compared with other machine learning approaches, such as prediction and classification algorithms, where accuracy and precision recall are more straightforward to implement. However, it is worth noting that RL has a distinct advantage over other machine learning approaches, that one can choose which outcome to optimize by specifying the reward function. This provides an opportunity to involve patient preferences and shared decision making. This becomes more relevant when learned policies change depending on the reward function. For example, an RL algorithm that optimizes survival may recommend a different set of treatments versus an RL algorithm that optimizes neurologic outcome. In such situations, patient preference is elicited to guide the choice of the RL algorithm.

RL has the potential to offer considerable advantages in supporting the decision making of physicians. However, certain key issues need to be addressed, such as clinical implementation, ethics, and medico-legal limitations in health care delivery [78]. In fact, any machine learning model would need to address these limitations carefully to serve as truly effective tools. In clinical practice, the RL models need to be refined iteratively throughout the time to include newly generated data from electronic health systems in hospitals, and the model must produce robust results for physicians to interpret and understand. Besides, patients' understanding and willingness to use the RL model as a supporting tool in their care would be another important consideration. Another important ethical consideration would be the liability in case of medical error when the RL model recommendation differs from the physician. It has an impact on the autonomy of both the physician and patient. The problem of medical error works in both ways when there is a poor outcome: (1) if the physician follows the RL model recommendation, can the clinician then blame the model and the personnel who maintain the model; (2) if the clinician does not follow the RL model recommendation, can the clinician then be said to have made the wrong decision and be penalized.

Possible directions for future work include (1) modeling the RL environment as a partially observable MDP, in which observations from the data are mapped to some state space that truly represents patients' underlying well-being; (2) extending the action space to be continuous, suggesting more precise and practical treatment recommendations to physicians; and (3) improving the interpretability of the RL models so that physicians can have more confidence in accepting the model results. With further efforts to tackle these challenges, RL methods could play a crucial role in helping to inform patient-specific decisions in critical care.

Conclusions

In this comprehensive review, we synthesized data from 21 articles on the use of RL to process or analyze retrospective data from ICU patients. With the improvement of data collection and advancement in reinforcement learning technologies, we see great potential in RL-based decision support systems to optimize treatment recommendations for critical care.

Acknowledgments

SL was funded by the National University of Singapore Graduate School for Integrative Sciences and Engineering Scholarship. This research was supported by the National Research Foundation Singapore under its AI Singapore Programme (award no. AISG-GC-2019-002), the National University Health System joint grant (WBS R-608-000-199-733), and the National Medical Research Council health service research grant (HSRG-OC17nov004).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Introduction to reinforcement learning.
[PDF File (Adobe PDF File), 617 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Summary of study characteristics.
[PDF File (Adobe PDF File), 167 KB-Multimedia Appendix 2]

References

1. Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Stat Med 2012 Jul 30;31(17):1887-1902 [FREE Full text] [doi: 10.1002/sim.4512] [Medline: 22438190]
2. Chen Z, Marple K, Salazar E, Gupta G, Tamil L. A physician advisory system for chronic heart failure management based on knowledge patterns. Theor Pract Log Prog 2016 Oct 14;16(5-6):604-618. [doi: 10.1017/S1471068416000429]
3. Hannes K, Leys M, Vermeire E, Aertgeerts B, Buntinx F, Depoorter A. Implementing evidence-based medicine in general practice: a focus group based study. BMC Fam Pract 2005 Sep 9;6:37 [FREE Full text] [doi: 10.1186/1471-2296-6-37] [Medline: 16153300]
4. Hutchinson A, Baker R. Making use of guidelines in clinical practice. Br Med J 1999 Oct 16;319(7216):1078 [FREE Full text] [doi: 10.1136/bmj.319.7216.1078] [Medline: 10521225]
5. James JT. A new, evidence-based estimate of patient harms associated with hospital care. J Patient Saf 2013 Sep;9(3):122-128. [doi: 10.1097/PTS.0b013e3182948a69] [Medline: 23860193]
6. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. Conf Proc IEEE Eng Med Biol Soc 2016 Aug;2016:2978-2981. [doi: 10.1109/EMBC.2016.7591355] [Medline: 28268938]
7. Zhang Z, Hong Y, Liu N. Scientific evidence underlying the recommendations of critical care clinical practice guidelines: a lack of high level evidence. Intensive Care Med 2018 Jul;44(7):1189-1191. [doi: 10.1007/s00134-018-5142-8] [Medline: 29564478]
8. Laffey JG, Kavanagh BP. Negative trials in critical care: why most research is probably wrong. Lancet Respir Med 2018 Sep;6(9):659-660. [doi: 10.1016/S2213-2600(18)30279-0] [Medline: 30061048]
9. Burke AE, Thaler KM, Geva M, Adiri Y. Feasibility and acceptability of home use of a smartphone-based urine testing application among women in prenatal care. Am J Obstet Gynecol 2019 Nov;221(5):527-528. [doi: 10.1016/j.ajog.2019.06.015] [Medline: 31300161]
10. Laserson J, Lantsman CD, Cohen-Sfady M, Tamir I, Goz E, Brestel C, et al. TextRay: Mining Clinical Reports to Gain a Broad Understanding of Chest X-Rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018 Presented at: MICCAI'18; September 16-20, 2018; Granada, Spain. [doi: 10.1007/978-3-030-00934-2_62]
11. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. JMLR Workshop Conf Proc 2016 Aug;56:301-318 [FREE Full text] [Medline: 28286600]
12. Fan J, Wang J, Chen Z, Hu C, Zhang Z, Hu W. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. Med Phys 2019 Jan;46(1):370-381. [doi: 10.1002/mp.13271] [Medline: 30383300]
13. Dagan N, Elnekave E, Barda N, Bregman-Amitai O, Bar A, Orlovsky M, et al. Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. Nat Med 2020 Jan;26(1):77-82. [doi: 10.1038/s41591-019-0720-z] [Medline: 31932801]

14. Watanabe AT, Lim V, Vu HX, Chim R, Weise E, Liu J, et al. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. J Digit Imaging 2019 Aug;32(4):625-637 [FREE Full text] [doi: 10.1007/s10278-019-00192-5] [Medline: 31011956]
15. Montague PR. Reinforcement learning: an introduction, by Sutton, RS and Barto, AG. Trends Cogn Sci 1999 Sep;3(9):360. [doi: 10.1016/S1364-6613(99)01331-5]
16. Kiumarsi B, Vamvoudakis KG, Modares H, Lewis FL. Optimal and autonomous control using reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst 2018 Jun;29(6):2042-2062. [doi: 10.1109/TNNLS.2017.2773458] [Medline: 29771662]
17. Howard RA. Dynamic Programming and Markov Process. New York, USA: MIT Press and Wiley; 1960.
18. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing atari with deep reinforcement learning. arXiv preprint 2013:- epub ahead of print(1312.5602) [FREE Full text]
19. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016 Jan 28;529(7587):484-489. [doi: 10.1038/nature16961] [Medline: 26819042]
20. Ng A, Coates A, Diel M, Ganapathi V, Schulte J, Tse B, et al. Autonomous inverted autonomous helicopter flight via reinforcement learning. In: Experimental Robotics IX. New York, USA: Springer; 2006:363-372.
21. 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 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
22. Riedmiller M. Neural Fitted Q Iteration – First Experiences with a Data Efficient Neural Reinforcement Learning Method. In: Proceedings of the European Conference on Machine Learning. 2005 Presented at: ECML'05; October 3-7, 2005; Porto, Portugal URL: https://doi.org/10.1007/11564096_32 [doi: 10.1007/11564096_32]
23. van Hasselt H, Guez A, Silver D. Deep Reinforcement Learning With Double Q-learning. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016 Presented at: AAAI'16; February 12-17, 2016; Phoenix, Arizona, USA. [doi: 10.5555/3016100.3016191]
24. Mnih V, Puigdomenech A, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. Arxiv 2016:- epub ahead of print(1602.01783) [FREE Full text]
25. Doya K, Samejima K, Katagiri K, Kawato M. Multiple model-based reinforcement learning. Neural Comput 2002 Jun;14(6):1347-1369. [doi: 10.1162/089976602753712972] [Medline: 12020450]
26. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: a brief survey. IEEE Signal Process Mag 2017 Nov;34(6):26-38. [doi: 10.1109/MSP.2017.2743240]
27. Wiering M, van Otterlo M, editors. Reinforcement Learning: State-of-the-Art. Berlin, Heidelberg: Springer-Verlag; 2012.
28. Siqi L. Reinforcement-Learning. GitHub. URL: https://github.com/nus-mornin-lab/Reinforcement-Learning [accessed 2020-01-01]
29. Borera EC, Moore BL, Doufas AG, Pyeatt LD. An Adaptive Neural Network Filter for Improved Patient State Estimation in Closed-Loop Anesthesia Control. In: 23rd International Conference on Tools with Artificial Intelligence. 2011 Presented at: ICTAI'11; November 7-9, 2011; Boca Raton, FL, USA. [doi: 10.1109/ictai.2011.15]
30. Padmanabhan R, Meskin N, Haddad WM. Closed-loop Control of Anesthesia and Mean Arterial Pressure Using Reinforcement Learning. In: Symposium on Adaptive Dynamic Programming and Reinforcement Learning. 2014 Presented at: ADPRL'14; December 9-12, 2014; Orlando, FL, USA. [doi: 10.1109/ADPRL.2014.7010644]
31. Padmanabhan R, Meskin N, Haddad WM. Reinforcement Learning-Based Control for Combined Infusion of Sedatives and Analgesics. In: 4th International Conference on Control, Decision and Information Technologies. 2017 Presented at: CoDIT'17; April 5-7, 2017; Barcelona, Spain. [doi: 10.1109/codit.2017.8102643]
32. Padmanabhan R, Meskin N, Haddad WM. Optimal adaptive control of drug dosing using integral reinforcement learning. Math Biosci 2019 Mar;309:131-142. [doi: 10.1016/j.mbs.2019.01.012] [Medline: 30735696]
33. Ghassemi MM, Alhanai T, Westover MB, Mark TG, Nemati S. Personalized Medication Dosing Using Volatile Data Streams. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. 2018 Presented at: AAAI'18; February 2-7, 2018; New Orleans, Louisiana, USA URL: https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/17234
34. Lin R, Stanley MD, Ghassemi MM, Nemati S. A deep deterministic policy gradient approach to medication dosing and surveillance in the ICU. Conf Proc IEEE Eng Med Biol Soc 2018 Jul;2018:4927-4931 [FREE Full text] [doi: 10.1109/EMBC.2018.8513203] [Medline: 30441448]
35. Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M. Deep reinforcement learning for sepsis treatment. arXiv 2017:- epub ahead of print(1711.09602) [FREE Full text]
36. Raghu A, Komorowski M, Celi L, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. arXiv 2017:- epub ahead of print(1705.08422) [FREE Full text]
37. Raghu A, Komorowski M, Singh S. Model-based reinforcement learning for sepsis treatment. arXiv 2018:- epub ahead of print-1811.09602 [FREE Full text]
38. Futoma J, Lin A, Sendak M, Bedoya A, Clement M, O'Brien C, et al. Learning to Treat Sepsis with Multi-Output Gaussian Process Deep Recurrent Q-Networks. OpenReview. 2018. URL: https://openreview.net/forum?id=SyxCqGbRZ [accessed 2020-06-10]

39. Peng X, Ding Y, Wihl D, Gottesman O, Komorowski M, Lehman LH, et al. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. AMIA Annu Symp Proc 2018;2018:887-896 [FREE Full text] [Medline: 30815131]
40. Lee D, Srinivasan S, Doshi-Velez F. Truly Batch Apprenticeship Learning with Deep Successor Features. 2019 Presented at: International Joint Conferences on Artificial Intelligence Organization; August 10-16, 2019; Macao, China. [doi: 10.24963/ijcai.2019/819]
41. Lopez-Martinez D, Eschenfeldt P, Ostvar S, Ingram M, Hur C, Picard R. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep Q networks. Conf Proc IEEE Eng Med Biol Soc 2019 Jul;2019:3960-3963. [doi: 10.1109/EMBC.2019.8857295] [Medline: 31946739]
42. Petersen B, Yang J, Grathwohl WS, Cockrell C, Santiago C, An G, et al. Precision medicine as a control problem: Using simulation and deep reinforcement learning to discover adaptive, personalized multi-cytokine therapy for sepsis. arXiv preprint. arXiv:1802.10440 2018 [FREE Full text]
43. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt B. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. Arxiv 2017:- epub ahead of print(1704.06300) [FREE Full text]
44. Cheng L, Prasad N, Engelhardt BE. An optimal policy for patient laboratory tests in intensive care units. Pac Symp Biocomput 2019;24:320-331 [FREE Full text] [Medline: 30864333]
45. 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 Apr 9;19(Suppl 2):57 [FREE Full text] [doi: 10.1186/s12911-019-0763-6] [Medline: 30961594]
46. Weng W, Gao M, He Z, Yan S, Szolovits P. Representation and reinforcement learning for personalized glycemic control in septic patients. arXiv 2017:- epub ahead of print(1712.00654) [FREE Full text]
47. Wang L, Zhang W, He X, Zha H. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018 Presented at: KDD'18; August 19-23, 2018; London, UK. [doi: 10.1145/3219819.3219961]
48. Manrai AK, Patel CJ, Ioannidis JP. In the era of precision medicine and big data, who is normal? J Am Med Assoc 2018 May 15;319(19):1981-1982. [doi: 10.1001/jama.2018.2009] [Medline: 29710130]
49. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
50. Ng A. CS294A Lecture Notes: Sparse Autoencoder. Stanford University. 2011. URL: https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf [accessed 2020-06-09]
51. Kane-Gill SL, Dasta JF, Buckley MS, Devabhakthuni S, Liu M, Cohen H, et al. Clinical practice guideline: safe medication use in the ICU. Crit Care Med 2017 Sep;45(9):e877-e915. [doi: 10.1097/CCM.0000000000002533] [Medline: 28816851]
52. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
53. Sinderby C, Breck J, Brander L. Bedside monitoring of diaphragm electrical activity during mechanical ventilation. In: Vincent JL, editor. Yearbook of Intensive Care and Emergency Medicine. New York, NY: Springer; 2009:385-393.
54. Patel SB, Kress JP. Sedation and analgesia in the mechanically ventilated patient. Am J Respir Crit Care Med 2012 Mar 1;185(5):486-497. [doi: 10.1164/rccm.201102-0273CI] [Medline: 22016443]
55. Cohen J, Shapiro M, Grozovski E, Singer P. Automatic tube compensation-assisted respiratory rate to tidal volume ratio improves the prediction of weaning outcome. Chest 2002 Sep;122(3):980-984. [doi: 10.1378/chest.122.3.980] [Medline: 12226043]
56. Kashyap R, Loftsgard TO. Clinicians role in reducing lab order frequency in ICU settings. J Perioper Crit Intens Care Nurs 2015;2(1):320-331 [FREE Full text] [doi: 10.4172/jpcic.1000112]
57. Osborne MA, Roberts SJ, Rogers A, Ramchurn SD, Jennings NR. Towards Real-Time Information Processing of Sensor Network Data Using Computationally Efficient Multi-output Gaussian Processes. In: International Conference on Information Processing in Sensor Networks. 2008 Presented at: IPSN'08; April 22-24, 2008; St Louis, MO, USA. [doi: 10.1109/ipsn.2008.25]
58. Kent JT. Information gain and a general measure of correlation. Biometrika 1983 Apr;70(1):163-173. [doi: 10.2307/2335954]
59. Liu X, Kruger P, Roberts MS. Optimizing drug dosing in the ICU. In: Yearbook of Intensive Care and Emergency Medicine. New York, USA: Springer; 2009.
60. Smith M. Monitoring and managing raised intracranial pressure after traumatic brain injury. In: Vincent JL, editor. Intensive Care Medicine. New York, NY: Springer; 2009:801-808.
61. Iannuzzi E, Iannuzzi M, Viola G, Sidro L, Cardinale A, Chiefari M. BIS - AAI and clinical measures during propofol target controlled infusion with Schnider's pharmacokinetic model. Minerva Anestesiol 2007;73(1-2):23-31 [FREE Full text] [Medline: 17115013]
62. Levi M. Emergency reversal of antithrombotic treatment. Intern Emerg Med 2009 Apr;4(2):137-145. [doi: 10.1007/s11739-008-0201-8] [Medline: 19002653]

63. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, et al. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 2011 May;39(5):952-960 [FREE Full text] [doi: 10.1097/CCM.0b013e31820a92c6] [Medline: 21283005]
64. Cohen J, Vincent J, Adhikari NK, Machado FR, Angus DC, Calandra T, et al. Sepsis: a roadmap for future research. Lancet Infect Dis 2015 May;15(5):581-614. [doi: 10.1016/S1473-3099(15)70112-X] [Medline: 25932591]
65. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). J Am Med Assoc 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
66. Jeter R, Josef C, Shashikumar S, Nemati S. Does the artificial intelligence clinician learn optimal treatment strategies for sepsis in intensive care? arXiv 2019:- epub ahead of print(1902.03271) [FREE Full text]
67. Lillicrap T, Hunt J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. arXiv 2015:- epub ahead of print(1509.02971) [FREE Full text]
68. An G. In silico experiments of existing and hypothetical cytokine-directed clinical trials using agent-based modeling. Crit Care Med 2004 Oct;32(10):2050-2060. [doi: 10.1097/01.ccm.0000139707.13729.7d] [Medline: 15483414]
69. Barr J, Fraser GL, Puntillo K, Ely EW, Gélinas C, Dasta JF, American College of Critical Care Medicine. Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. Crit Care Med 2013 Jan;41(1):263-306. [doi: 10.1097/CCM.0b013e3182783b72] [Medline: 23269131]
70. Yu G, Zhang F, Tang S, Lai M, Su R, Gong Z. Continuous infusion versus intermittent bolus dosing of morphine: a comparison of analgesia, tolerance, and subsequent voluntary morphine intake. J Psychiatr Res 2014 Dec;59:161-166. [doi: 10.1016/j.jpsychires.2014.08.009] [Medline: 25193460]
71. Rutter PC, Murphy F, Dudley HA. Morphine: controlled trial of different methods of administration for postoperative pain relief. Br Med J 1980 Jan 5;280(6206):12-13 [FREE Full text] [doi: 10.1136/bmj.280.6206.12] [Medline: 6986940]
72. Nagaraj SB, McClain LM, Zhou DW, Biswal S, Rosenthal ES, Purdon PL, et al. Automatic classification of sedation levels in ICU patients using heart rate variability. Critical Care Medicine 2016;44(9):e782-e789. [doi: 10.1097/ccm.0000000000001708] [Medline: 27035240]
73. Carrasco G. Instruments for monitoring intensive care unit sedation. Crit Care 2000;4(4):217-225 [FREE Full text] [doi: 10.1186/cc697] [Medline: 11094504]
74. Lewis SR, Pritchard MW, Fawcett LJ, Punjasawadwong Y. Bispectral index for improving intraoperative awareness and early postoperative recovery in adults. Cochrane Database Syst Rev 2019 Sep 26;9(6):CD003843 [FREE Full text] [doi: 10.1002/14651858.CD003843.pub4] [Medline: 31557307]
75. Williams JD, Young S. Partially observable Markov decision processes for spoken dialog systems. Comput Speech Lang 2007 Apr;21(2):393-422. [doi: 10.1016/j.csl.2006.06.008]
76. Maei H, Szepesv´ari C, Bhatnagar S, Sutton R. Toward Off-Policy Learning Control with Function Approximation. In: The 27th International Conference on Machine Learning. 2010 Presented at: ICML'10; June 21-24, 2010; Haifa, Israel URL: https://icml.cc/Conferences/2010/papers/627.pdf
77. McShea M, Holl R, Badawi O, Riker RR, Silfen E. The eICU research institute - a collaboration between industry, health-care providers, and academia. IEEE Eng Med Biol Mag 2010;29(2):18-25. [doi: 10.1109/MEMB.2009.935720] [Medline: 20659837]
78. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4]

Abbreviations

AI: artificial intelligence
aPTT: activated partial thromboplastin time
BIS: bispectral index
BUN: blood urea nitrogen
DNN: deep neural network
DQN: deep Q network
eRI: eICU Research Institute
FQI: fitted-Q-Iteration
ICU: intensive care unit
IV: intravenous
MAP: mean arterial pressure
MDP: Markov decision process
MIMIC III: Medical Information Mart for Intensive Care III
MV: mechanical ventilation
RCT: randomized controlled trial
RL: reinforcement learning

WBC: white blood cell

Edited by G Eysenbach; submitted 28.02.20; peer-reviewed by A Hallawa, Z Zhang, D Maslove; comments to author 08.04.20; revised version received 05.05.20; accepted 13.05.20; published 20.07.20

Please cite as:
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
URL: https://www.jmir.org/2020/7/e18477
doi: 10.2196/18477
PMID: