Surgery xxx (2020) 1–4

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

Reinforcement learning in surgery

Shounak Datta, PhD[a,e], Yanjun Li, MS[b,e], Matthew M. Ruppert, BS[a,e], Yuanfang Ren, PhD[a,e], Benjamin Shickel, MS[c,e], Tezcan Ozrazgat-Baslanti, PhD[a,e], Parisa Rashidi, PhD[d,e], Azra Bihorac, MD, MS[a,e,*]

[a] Department of Medicine, College of Medicine, University of Florida, Gainesville, FL
[b] NSF Center for Big Learning, University of Florida, Gainesville, FL
[c] Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL
[d] Department of Biomedical Engineering, College of Medicine, University of Florida, Gainesville, FL
[e] Precision and Intelligent Systems in Medicine (PRISMA^P), University of Florida, Gainesville, FL

ARTICLE INFO

Article history:
Accepted 27 November 2020
Available online xxx

ABSTRACT

Patients and physicians make essential decisions regarding diagnostic and therapeutic interventions. These actions should be performed or deferred under time constraints and uncertainty regarding patients' diagnoses and predicted response to treatment. This may lead to cognitive and judgment errors. Reinforcement learning is a subfield of machine learning that identifies a sequence of actions to increase the probability of achieving a predetermined goal. Reinforcement learning has the potential to assist in surgical decision making by recommending actions at predefined intervals and its ability to utilize complex input data, including text, image, and temporal data, in the decision-making process. The algorithm mimics a human trial-and-error learning process to calculate optimum recommendation policies. The article provides insight regarding challenges in the development and application of reinforcement learning in the medical field, with an emphasis on surgical decision making. The review focuses on challenges in formulating reward function describing the ultimate goal and determination of patient states derived from electronic health records, along with the lack of resources to simulate the potential benefits of suggested actions in response to changing physiological states during and after surgery. Although clinical implementation would require secure, interoperable, livestreaming electronic health record data for use by virtual model, development and validation of personalized reinforcement learning models in surgery can contribute to improving care by helping patients and clinicians make better decisions.

Introduction

Patients and physicians make essential decisions regarding diagnostic and therapeutic interventions. These actions should be performed or deferred under time constraints and uncertainty regarding patients' diagnoses and predicted response to treatment. This may lead to cognitive and judgment errors. The uncertainty regarding patients' diagnoses and predicted response to treatment may lead to cognitive and judgment errors. Reinforcement learning is a subfield of machine learning that identifies a sequence of actions to increase the probability of achieving a predetermined goal. Reinforcement learning has the potential to assist in surgical decision-making. The trail-and-error learning approach recommends specific actions at predefined intervals and its ability to utilize complex input data, including text, image, and temporal data, in the decision-making process.[1] This review seeks to describe the challenges in development and application of reinforcement learning in health care, and in surgery in particular (Fig 1), for the applications themselves have previously been summarized by Yu et al,[2] Loftus et al,[1] and Liu.[3]

Reinforcement Learning

Reinforcement learning (RL) is a subfield of machine learning (ML) that identifies a sequence of actions to increase the probability of achieving a predetermined goal. It is a technique for developing powerful solutions in a variety of health care domains, where diagnosing decisions or treatment regimens are usually characterized by a sequential decision-making procedure.[2] A RL problem is solved through a trial-and-error learning process, emulating human learning behavior. A RL agent (part of algorithm suggesting actions) interacts with an

Shounak Datta, Yanjun Li, and Matthew M. Ruppert contributed equally to the manuscript. T.O.B., P.R., and A.B. served as senior authors.

* Reprint requests: Azra Bihorac, Department of Medicine, Precision and Intelligent Systems in Medicine (Prisma^P), Division of Nephrology, Hypertension, and Renal Transplantation, PO Box 100224, Gainesville, FL 32610-0224.
E-mail address: abihorac@ufl.edu (A. Bihorac);
Twitter: @azrabihorac

https://doi.org/10.1016/j.surg.2020.11.040
0039-6060/

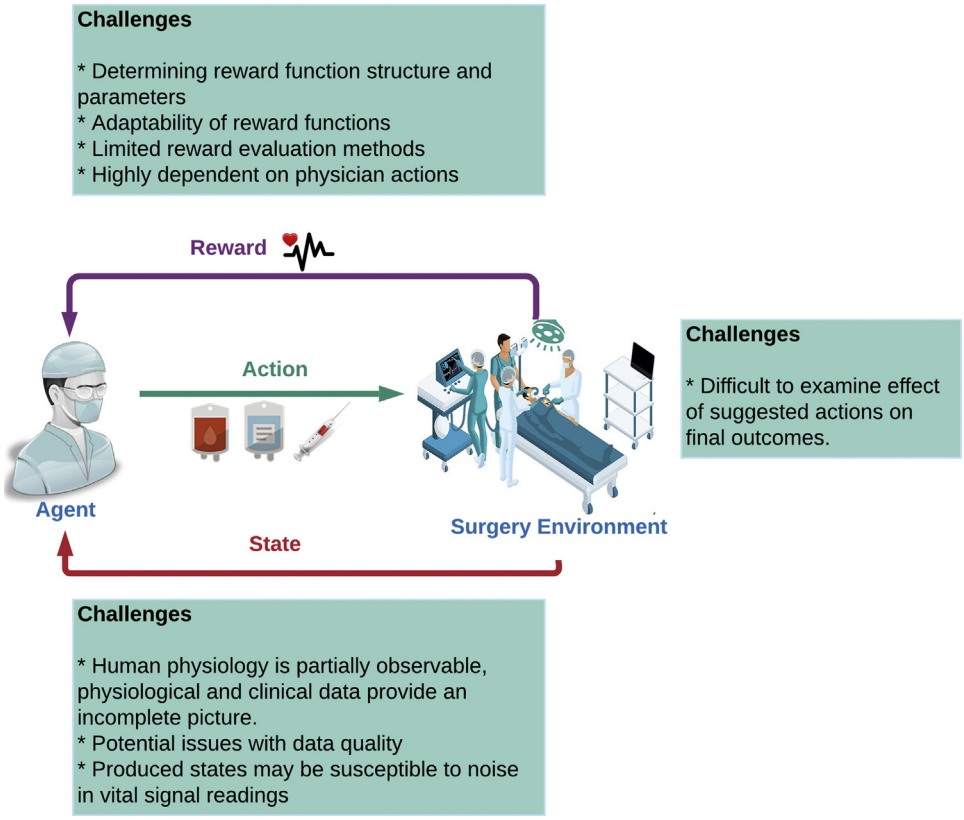

Fig. 1. Reinforcement learning framework and challenges in development of reinforcement learning models.

environment (patient/surgical session observed from electronic health records) to maximize the cumulative reward resulting from its actions. Generally, RL problems are modeled and solved using a Markov decision process (MDP), guided by Bellman's equation.[4] There are 4 components: (1) a state that represents the environment at each time step; (2) an action the agent takes at each time step that influences the next state; (3) a transition probability that provides an estimate for reaching different subsequent states, which reflects the environment in which an agent interacts; and (4) a reward function, which is the observed feedback given a state-action pair (Fig 1).

Applications of reinforcement learning in health care

In recent times, the reinforcement learning approach is being studied in health care to produce optimum policies to suggest interventions and recommend actions to omit human-level bias and errors. It has the ability to mimic a human-like learning approach and can use electronic health records to develop treatment policies, intervention suggestion systems, and action recommendation systems, and it has the potential to improve care by helping clinicians make better decisions. The robustness of recent RL algorithms help the developed systems adapt to even sudden changes in patient physical states. Yu et al[2] discussed the broad applications of RL techniques in health care domains. One of the main applications is in dynamic treatment regimens, which provide sequential clinical decision-making (eg, drug dosage, intervention time, or treatment type) for individual patients with long-term care. Komorowski et al[5] used 48 variables to describe patient state, including demographics, comorbidities, vital signs, and laboratory values. They aggregated data for each patient every 4 hours, clustered each description into 750 discrete, mutually exclusive groups to create a state space, and applied a policy iteration RL algorithm to learn the optimal dosing policy for intravenous fluid and vasopressors that will maximize the 90-day survival probability. Prasad et al[6] extracted records of patients with ventilator support every 10 min and applied the RL-based fitted Q-Iteration (FQI) method to optimize the mechanical ventilation (MV) and sedation weaning time. Another major application is the automation of medical diagnoses, where a diagnosis is formulated through a sequential decision-making process. Ling et al[7] proposed a novel approach for clinical diagnosis inferencing that applied deep Q-learning (DQN) to learn the optimal policy to obtain a final diagnosis through iterative search through candidate diagnoses from an external resource (Wikipedia and Mayo Clinic). Other applications include health resource scheduling and allocation,[8] optimal process control,[9] and drug discovery and development[10] (Table I).

Knowledge Gaps

Challenges in reward formulation

As described in detail by Yu et al,[2] formulating the reward function for reinforcement learning is one of the most challenging aspects of applying such models in health care, especially for intraoperative applications. Reward functions are used to generate a reward score R_{t+1} for taking an action a_t at state S_t at a given timepoint t to transition to state S_{t+1}. Formulation of a reward function requires a comprehensive understanding of both short- and long-term goals, as well as how the states and environment are defined. Most RL algorithms are tested and evaluated in the context of gaming or robotics, where such information is readily available. However, in health care applications, the reward functions must be generated by attributing numeric values that indicate the degree of benefit or harm from state changes derived from vital signs, clinical notes, clinical images, laboratory data, demographic, and socioeconomic information.

For example, Komorowski et al[5] trained a fluid and vasopressor dosing algorithm for sepsis in ICU patients solely on the basis of

Table I
Summary of relevant studies addressing RL challenges in health care application

Authors	Population	Sample size	Algorithms	Study Objective	Major Finding
Komorowski et al (2018)[5]	Septic patients in ICU	96,000	SARSA algorithm	Reduce 90-day mortality by suggesting appropriate IV fluid and vasopressor dose	Developed system was able to propose IV and vasopressor dosage for every 4-h window utilizing clustered physiological states
Prasad et al (2017)[6]	ICU patients successfully discharged after stay	6,883	Deep Q Network	Develop optimum weights to guide short-term reward function design	Produced a more optimal way of generating short-term rewards weights to develop improved ICU action policy
Ling et al (2017)[7]	Not applicable	Not applicable	Deep Q-network	Infer the patient's diagnoses from the clinical narrative and external knowledge	Formulated the process of a differential diagnosis as a reinforcement learning problem and demonstrated the effectiveness in determining the correct diagnosis compared to various nonreinforcement learning-based methods
Huang et al (2011)[8]	Not applicable	Not applicable	Q-learning	Optimize resource allocation in the business	Proposed approach that outperformed some heuristic or hand-coded strategies
Nguyen et al (2019)[9]	Not applicable	Not applicable	Deep Reinforcement Learning	Learn tensioning policies for surgical soft tissue cutting tasks	Proposed tensioning policy outperformed the state-of-the-art method with respect to both accuracy and reliability
Popova et al (2018)[10]	Not applicable	Not applicable	Deep Neural Network combined with Deep Reinforcement Learning	Design targeted chemical libraries of compounds with desired properties	Formulated chemical library design as a reinforcement learning problem and demonstrated the novelty and synthetic accessibility of generated chemicals
Dai et al (2020)[11]	Outpatients	Not mentioned	Deep Neural Network combined with Deep Reinforcement Learning	Develop robust treatment suggestion method using imaging information as states	Formulated a reward function as based on the l_2-norm distance between previous and current health conditions (using image data)
Yu et al (2020)[12]	ICU patients	8,600	Supervised Actor Critic algorithm	Develop optimum policy of controlling patient mechanical ventilation and sedative dosage during ICU stays	Developed a system utilizing short-term goals to predict the necessity of mechanical ventilation and appropriate sedative dose

ICU, intensive care unit; *SARSA*, state-action-reward-state-action.

minimizing the probability of 90-day mortality, ignoring any short-term indicators of health such as blood pressure or volume status. Although patient mortality increased as the actions of the physician and agent became more disparate, it is unclear if reward functions based exclusively on death are effective in such high mortality environments. On average, the algorithm suggested higher doses of vasopressors and IV fluids. Also, mortality was lowest when physician and model recommendation matched.

Dai et al[11] developed a more sophisticated approach to state representation and reward policy formulation by creating a 9-dimensional state representation using multiple deep neural networks and a reward function based on the squared distance between h^* and h, where h is the target health state and h^* is the resulted state from a simulation designed using deep neural networks. Although this approach holds promise in creating more generalizable state definitions and reward functions, the authors reported limited model optimization, resulting in unreliable treatment suggestions.

Yu et al[12] developed a supervised version of RL using an actor-critic approach. The actor-critic approach uses an actor to suggest best action (policy optimization), and a critic evaluates the action qualities (computing quality of suggested action).[13] Their aim was to train an algorithm to recognize when MV is warranted in ICU patients and recommend the optimal dose of propofol to keep the patient stably sedated while on MV. They used 13 vital signs along with age and weight to define the state and used only short-term changes in vital sign stability as the basis for the reward function.

In an effort to better define best practices for short-term reward function formulation in RL problems, Prasad et al[6] studied strategies for deriving reward weights that were better tailored to a given context. They reported a larger effective sample size (owing to their approach, which requires less data to converge compared with other techniques) after their optimized reward function approach was applied; however, there is a lack of knowledge as to how overall RL performance is improved through such tactics. Notably, their approach was highly dependent on the level of "correctness" of the physicians' actions used in the training.

Challenges in patient state determination

The first step to build a reinforcement learning system in surgery is to define the states, where each state is a complete description of a patient's physiological status. It is crucial to collect and summarize the pertinent health information for each patient state representation. This summarized information should be organized or preprocessed into a concise and manageable form to train the learning agent effectively and efficiently.

A majority of the current work leverages medical data that may include static traits of patients (eg, demographics such as age, sex, ethnicity, comorbidities), longitudinal measurements (eg, vital signs, laboratory values, physiological, pathologic),[3] and/or medical images[11] This raw data is then transformed into a uniform high-dimensional vector as the final state representation.[2] using pre-defined discretization methods or trainable methods including but not limited to linear models and deep neural networks. For example, Komorowski et al[5] consolidated thousands of combinations of 48 variables into 750 discrete mutually exclusive states.

Muddling this process are the data quality, inconsistency, noise, and missingness associated with electronic health records. Although numerous methods have been proposed to solve these problems, It is unclear how robust the current state formulation methods are to such problems, given the effect underlying noise and bias has on the formulation of patient states. Such an understanding is critical to building a successful reinforcement learning system in the surgery.

More importantly, most of the current work uses a MDP to model the patient states and trajectories. In current medical practice a patient's physiological state is approximated using readily

measurable or observable properties such as blood pressure; however, the underlying physiology that dictates the value of these properties is unobservable. Therefore, a partially observable MDP (POMDP) approach may be superior because it can theoretically use unobserved relationships in the determination of patient states.

Challenges in modeling physiological response to agent actions

The complexity of human physiology makes the training and implementation of reinforcement learning algorithms for surgical decision-making very difficult, with many unobservable variables that often get ignored. The dynamic mechanisms in which the human body responds to stimuli are still not completely understood, making it difficult to model since such responses often have systematic components that have varying effects on different parts of the body.[14]

Technology gaps

Current ML implementations in health care are based primarily on a centralized model in which data is aggregated and stored in a central environment, where it is then used for the training and implementation of chosen ML algorithms. Despite massive investments in infrastructure, this approach is still far from delivering true real-time execution due to bandwidth limitations in streaming clinical data for simultaneous processing of hundreds or thousands of patients. In addition to the technical costs and limitations, there are privacy and security concerns due to the constant transmission and aggregation of protected health information. An alternative approach, which has been gaining significant traction due to its privacy first focus, is on device ML. On device ML has many advantages including reducing network congestion, reducing execution time, and better protection of protected health information. On device ML has further been strengthened through a technique known as federated learning, first proposed by Google in 2016,[15] that allows for on device ML algorithms to share and aggregate knowledge without the sharing and aggregation of the underlying data.

Future directions

Detailed study of reward function design is critical for properly guiding the agent toward the desired outcome in a given environment. The algorithm mimics the human learning approach of trial and error, an important feature for developing artificial decision-making algorithms. The goal of this algorithm is to present improvement in clinical decision-making processes. A fundamental flaw in much of the work dedicated to addressing this issue is the assumption of physicians being the gold standard to measure the correctness of an agent's actions. One approach to validate the "correctness" of a physician's actions would be to randomly sample states and the actions taken based on those states from a pool of patients that a physician treated. The actions of the physician in these scenarios would then be compared against the actions proposed by a panel of subject matter experts. Although this approach may be superior in some respects to the reliance on a single physician for the determination of the most "correct" action, it still fails to address aspects of medicine that are still intangible to machines such as the impression that a physician has of a patient and the art of medicine itself. Patients are highly individual in both personality and physiology eliminating the notion of a one-size-fits-all approach to clinical decision-making, which necessitates a dynamic approach bespoke to each patient.[16] Given the complex, high-stakes, and often uncertain nature of surgical decision-making, a collaborative approach to decision making is often warranted where the all stakeholders (physicians, other health care team members, patients, and patients' families) can collectively design a plan that improves patient satisfaction and may reduce the costs associated with undesired treatments.[16] Future implementations of reinforcement learning in surgical settings should incorporate dynamic reward functions to accept input from both the patient and all members of a perioperative care team. Such collaborative reward functions can balance the risk aversion of individual patients and surgeons with the expected benefits and postoperative care trajectories highlighted by other team members. By giving the patient increased control over their own algorithm-influenced clinical care, collaborative and dynamic reward functions have the potential to increase overall patient satisfaction. Appropriate use of reinforcement learning in health care and in surgery may improve care by helping patients and clinicians to make better decisions and have better outcomes.

Funding/Support

A.B., T.O.B., and P.R. were supported by R01 GM110240 from the National Institute of General Medical Sciences. A.B. and T.O.B. were supported by Sepsis and Critical Illness Research Center Award P50 GM-111152 from the National Institute of General Medical Sciences. P.R. was supported by the NSF CAREER 1750192 and NIH/NIBIB 1R21EB027344 grants. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. T.O.B., P.R., and A.B. served as senior authors.

Conflict of interest/Disclosure

The authors report no conflict of interest.

References

1. Loftus TJ, Filiberto AC, Li Y, Balch J, Cook AC, Tighe PJ, et al. Decision analysis and reinforcement learning in surgical decision-making. *Surgery*. 2020;168:253–266.
2. Yu C, Liu J, Nemati S. Reinforcement learning in healthcare: a survey. arXiv preprint 2019;arXiv:1908.08796
3. 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:e18477.
4. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT Press; 1998.
5. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Med*. 2018;24:1716–1720.
6. Prasad N, Cheng L-F, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv preprint. 2017;arXiv:1704.06300.
7. Ling Y, Hasan SA, Datla V, Qadir A, Lee K, Liu J, et al. Learning to Diagnose: Assimilating Clinical Narratives using Deep Reinforcement Learning. *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. 2017;1(Long Papers):895–905.
8. Huang Z, van der Aalst WM, Lu X, Duan H. Reinforcement learning based resource allocation in business process management. *Knowl Eng*. 2011;70:127–145.
9. Nguyen ND, Nguyen T, Nahavandi S, Bhatti A, Guest G. Manipulating Soft Tissues by Deep Reinforcement Learning for Autonomous Robotic Surgery. *2019 IEEE International Systems Conference (SysCon)*. 2019:1–7.
10. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv*. 2018;4:eaap7885.
11. Yu C, Ren G, Dong Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med Inform Decis Mak*. 2020;20:124. https://doi.org/10.1186/s12911-020-1120-5.
12. Song R, Lewis FL, Wei Q, Zhang H. Off-Policy Actor-Critic Structure for Optimal Control of Unknown Systems With Disturbances. *IEEE Transactions on Cybernetics*. 2016;46:1041–1050.
13. Dai Y, Wang G, Muhammad K, Liu S. A closed-loop healthcare processing approach based on deep reinforcement learning. *Multimed Tools Appl*. 2020. https://doi.org/10.1007/s11042-020-08896-5.
14. Karin O, Swisa A, Glaser B, Dor Y, Alon U. Dynamical compensation in physiological circuits. *Mol Syst Biol*. 2016;12:886.
15. Konečný J, McMahan H, Ramage D, Richtárik P. Federated optimization: distributed machine learning for on-device intelligence. ArXiv preprint. 2016;arXiv:1610.02527
16. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. *JAMA Surg*. 2020;155:148–158.