The promises and pitfalls of reinforcement learning in healthcare

Omer Gottesman[*1], Fredrik Johansson[*2], Matthieu Komorowski[3,4], Aldo Faisal[5], David Sontag[2], Finale Doshi-Velez[1], Leo Anthony Celi[3,6,7]

[*] These authors contributed equally to the preparation of the manuscript

[1] Paulson School of Engineering and Applied Sciences, Harvard University

[2] Institute for Medical Engineering and Science, MIT

[3] Laboratory for Computational Physiology, Harvard-MIT Health Sciences & Technology, MIT

[4] Department of Surgery and Cancer, Faculty of Medicine, Imperial College London

[5] Department of Bioengineering, Imperial College London

[6] Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center

[7] MIT Critical Data

Abstract

In this Comment, we provide guidelines for reinforcement learning for patient treatment decisions that we hope will accelerate the rate at which observational cohorts can inform healthcare practice in a safe, risk-conscious manner.

From sepsis warning systems to identifying subtle disease signals in medical images, artificial intelligence (AI) is poised to transform healthcare for the better[1]. However, AI is not a panacea, and if used improperly, these systems can replicate our bad practices rather than improve them.

Reinforcement Learning (RL) is a subfield of AI that provides tools to optimize sequences of decisions for long-term outcomes. For example, faced with a patient with sepsis, the intensivist must decide if and when to initiate and adjust treatments such as antibiotics, intravenous fluids, vasopressor agents, and mechanical ventilation. Each choice affects the patient's survival at the end of the hospital stay, quality of life upon recovery, and so on. While the RL approaches used to optimize treatment sequences vary, they all fall into a common framework. RL algorithms take as input sequences of interactions (called histories) between the decision-maker and their environment. At every decision-point, the RL algorithm chooses an action according to its policy and receives new observations and immediate outcomes (often called rewards).

In the context of healthcare, RL has been applied to optimizing anti-retroviral therapy in HIV[2], tailoring anti-epilepsy drugs for seizure control[3], and determining the best approach to managing sepsis[4]. In contrast to more common uses of AI such as one-time predictions, the output (decisions) of an RL system affects both the patient's future health and future treatment options[5]. As a result, long-term effects are harder to estimate (Figure 1a).

To illustrate the potential pitfalls in reinforcement learning, we use the example of sepsis management, for which there remains wide variability in the way clinicians make decisions. In the context of sepsis, a history may include a patient's vital signs and laboratory tests. The actions are all the treatments available to the clinician, including medications and interventions. The rewards require clinician input: they should represent the achievement of desirable tasks such as stabilization of vital signs or survival at the end of the stay. By weighing different

rewards, an RL-algorithm could be designed to target short-term outcomes, such as liberation from mechanical ventilation, or longer-term outcomes, such as prevention of permanent organ damage. Note that defining short-term goals is not straightforward, since ideal sepsis resuscitation targets remain elusive[6].

We discuss three key questions that should be considered when reading an RL study. These questions uncover limitations when making quantitative performance claims about RL-learned algorithms from observational data.

Is the AI given access to all the variables that influence decision making?

A clinician could not be expected to make good decisions about a patient's vasopressor medication dosing without knowing about the patient's co-morbid cardiac condition as well as what has transpired in the last 24 hours, and neither can an AI. To estimate the quality of a new treatment policy based on historical data, it is vital to take into account any information that was used by clinicians in their decision making—failing to do so may result in estimates that are confounded by spurious correlation. For example, severely sick septic patients may receive fluids earlier, yet have worse outcomes than healthier patients which is clearly a result of them being sicker in the first place. This difference in outcomes may lead an analysis that associates earlier fluid administration with worse outcomes if not properly adjusted for clinical context. Adjusting for confounding is challenging when validating the average treatment effect of a single decision[7]; this problem becomes significantly harder when decisions are made in sequence. It is thus important to be conscientious of possible confounding factors when reading an RL study even more so than for standard prediction studies, as the sequential nature of the problem could lead to confounding effects on the long as well as the short term.

Effective cohort size: How big was that big data, really?

When evaluating the quality of an RL algorithm retrospectively, the choice of the proposed treatment policy affects the effective sample size. This occurs because most approaches for evaluating RL policies from observational data weigh each patient's history based on whether the clinician decisions match the decisions of the policy proposed by the RL algorithm[8]. The reliability (variance) of the treatment quality estimate depends on the number of patient histories for which the proposed and observed treatment policies agree — a quantity known as the effective sample size. The possibilities for mismatch between the actual decision and the proposed decision grow with the number of decisions in the patient's history, and thus RL-based evaluations are especially prone to having small *effective* sample sizes (Figure 1b).

For example, we found that the effective sample size for a sepsis management policy on a cohort of 3855 patients was only a few dozens[9]. In general, the effective sample size will be larger if the learned policies are close to the clinician policies, suggesting that RL with observational data will be most reliable for refining existing practices rather than discovering new treatment approaches.

Will the AI behave prospectively as intended?

Even if the AI has access to all the important variables and the evaluation was perfect, errors in problem formulation or data processing can lead to poor decisions. Simplistic reward functions may neglect long-term effects for meaningless gains: for example, rewarding only blood pressure targets may result in an AI that causes long-term harm by excessive dosing of vasopressors. Errors in data recording or preprocessing may introduce errors in the reward signal, misleading the RL algorithm. Finally, the learned policy may not work well at a different hospital or even in the same hospital a year later if treatment standards shift.

Thus, it is essential to interrogate RL-learned policies to assess whether they will behave

prospectively as intended. An increasing body of work on interpretable machine learning enables such introspection[10].

Toward standard practice

Together, big data and RL provide unique opportunities for optimizing treatments in healthcare, especially those undertaken in sequence. However, to realize this potential, we must exercise caution and due diligence in their application and evaluation.

Figures

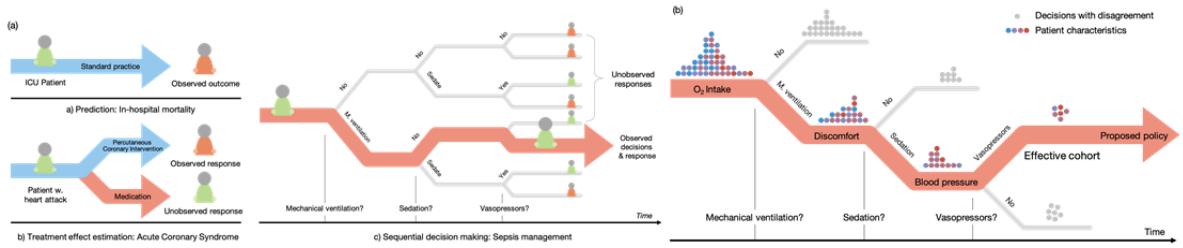


Figure 1. (a) Prediction, treatment effect estimation and sequential decision-making tasks. These tasks are progressively harder to solve based on observational data. In classical prediction tasks, only a single outcome for a patient is considered—the result of following standard practice without interventions from the analyst. Here, we use the common example of predicting 48h in-hospital mortality. In treatment effect estimation, we must also reason about what would happen under alternative unobserved interventions. Consider for example choosing between performing catheterization on a patient with cardiac arrest, or placing them on medication. To perform sequential decision making, such as for sepsis management, treatment effect estimation must be solved at a much grander scale—every possible combination of interventions could be considered to find an optimal treatment policy. **(b)** Effective sample size in off-policy evaluation. Each dot represents a single patient at each stage of treatment, its color indicating the patient's characteristics. The more decisions are performed in sequence, the likelier it is that a new policy disagrees with the one used to learn from. We illustrate disagreement by grayed out decision points. Using only samples for which the old policy agrees with the new results in a small effective sample size and a biased cohort, as illustrated by the difference in color distribution in the original and final cohort.

1 Obermeyer, Z. & Emanuel, E. J. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* **375**, 1216 (2016).

2 Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V. & Doshi-Velez, F. Combining Kernel and Model Based Learning for HIV Therapy Selection. *AMIA Summits on Translational Science Proceedings* **2017**, 239 (2017).

3 Guez, A., Vincent, R. D., Avoli, M. & Pineau, J. in *AAAI* 1671-1678 (2008).

4 Komorowski, M., Celi, L. A., Badawi, O., Faisal, A. & Gordon, A. The intensive care AI clinician learns optimal treatment strategies for sepsis. *Nature Medicine* (2018 (In press)).

5 Chakraborty, B. & Moodie, E. *Statistical methods for dynamic treatment regimes*. (Springer, 2013).

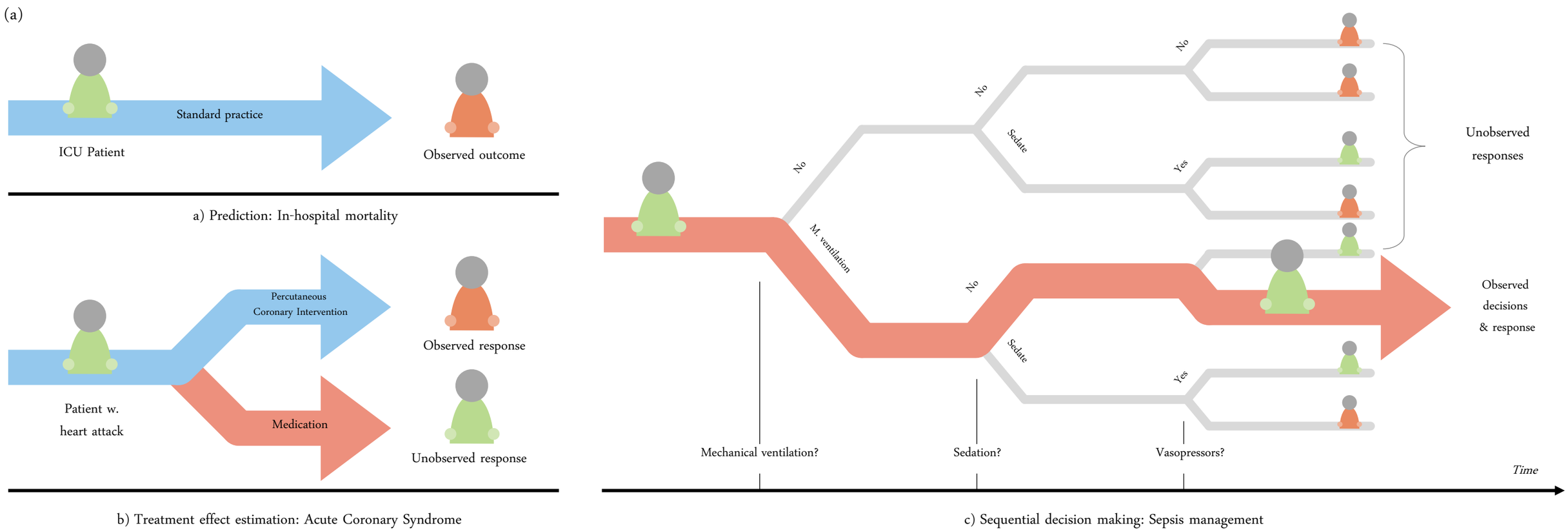
6 Simpson, N., Lamontagne, F. & Shankar-Hari, M. Septic shock resuscitation in the first hour. *Current opinion in critical care* **23**, 561-566 (2017).

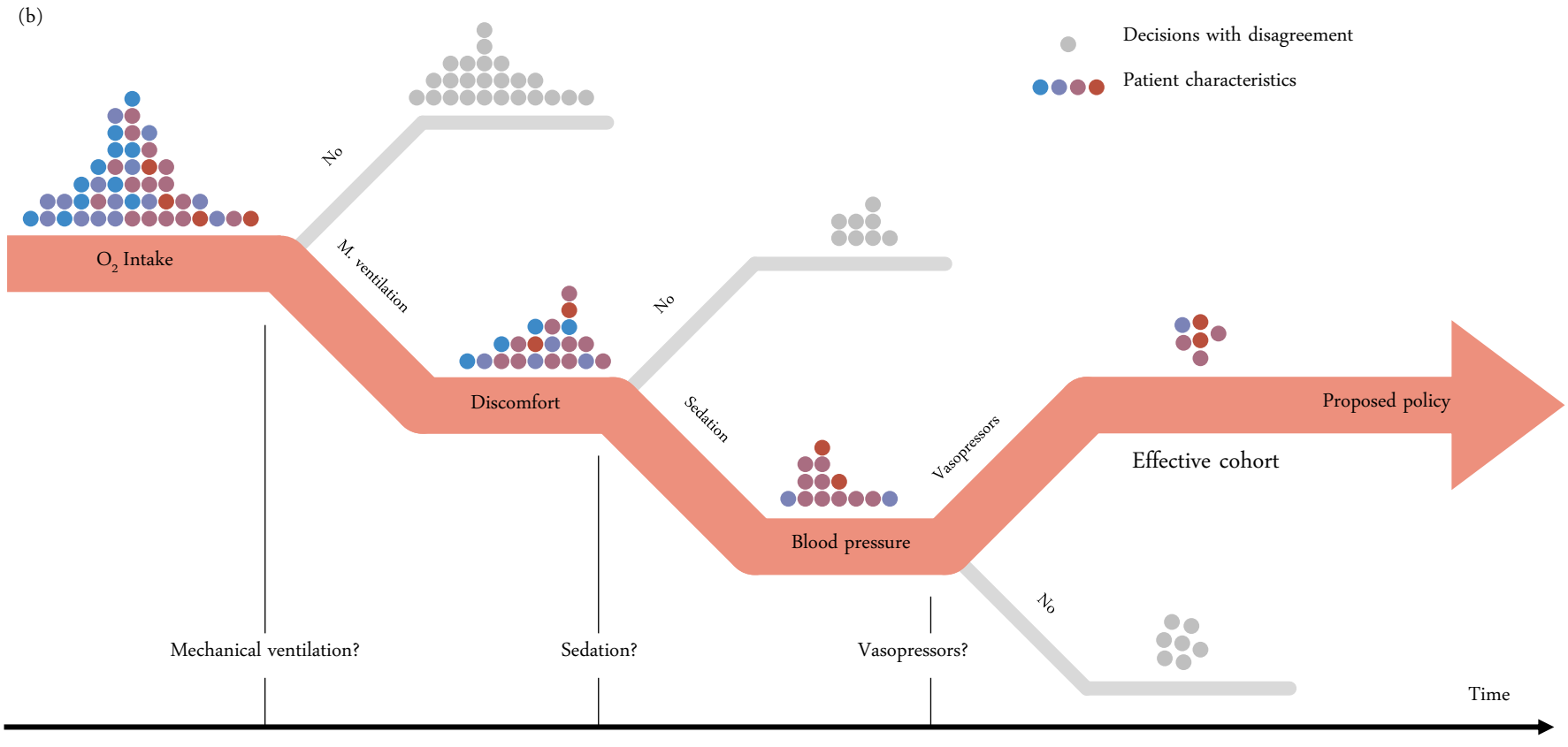
7 Johansson, F., Shalit, U. & Sontag, D. in *International Conference on Machine Learning* 3020-3029 (2016).

8 Precup, D., Sutton, R. S. & Singh, S. P. in *International Conference on Machine Learning* 759-766 (2000).

9 Gottesman, O. *et al.* Evaluating Reinforcement Learning Algorithms in Observational Health Settings. *arXiv preprint arXiv:1805.12298* (2018).

10 Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).


(a)
Standard practice
ICU Patient
Observed outcome
a) Prediction: In-hospital mortality
Percutaneous
Coronary Intervention
Medication
Patient w.
heart attack
Observed response
Unobserved response
b) Treatment effect estimation: Acute Coronary Syndrome
No
M. ventilation
No
Sedate
No
Yes
No
Sedate
Yes
Unobserved
responses
Observed
decisions
& response
Mechanical ventilation?
Sedation?
Vasopressors?
Time
c) Sequential decision making: Sepsis management


(b)
Decisions with disagreement
Patient characteristics
O_2 Intake
No
M. ventilation
Discomfort
No
Sedation
Blood pressure
Vasopressors
Proposed policy
Effective cohort
No
Mechanical ventilation?
Sedation?
Vasopressors?
Time