Guidelines for reinforcement learning in healthcare

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

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez and Leo Anthony Celi

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 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, the quality of the patient's life upon recovery, and so on. While the RL approaches currently 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 antiretroviral therapy in HIV[2], tailoring antiepilepsy drugs for seizure control[3], and determining the best approach to managing sepsis[4]. In contrast with more common uses of AI, such as one-time predictions, the output (or the decision) of a RL system affects both the patient's future health and future treatment options[5]. As a result, long-term effects are harder to estimate (Fig. 1).

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, a 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].

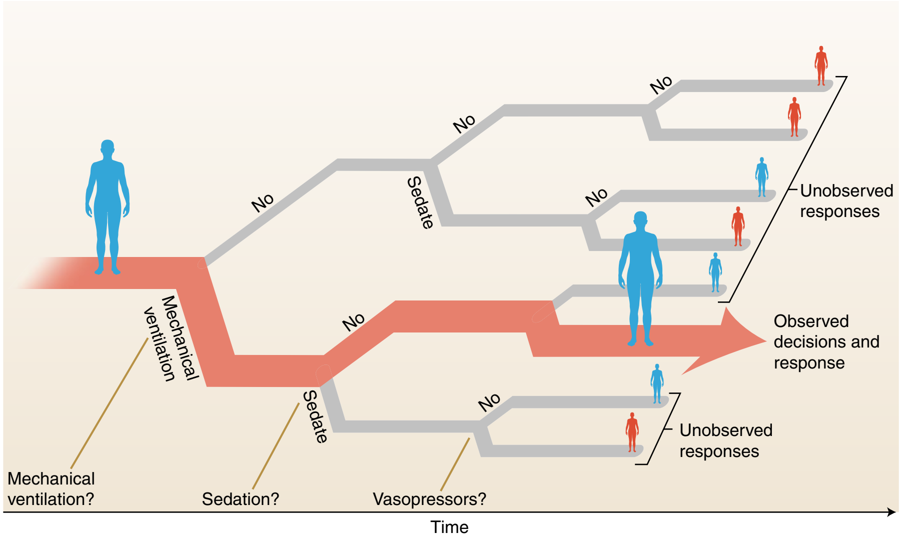

Fig. 1 | Sequential decision-making tasks. To perform sequential decision making, such as for sepsis management, treatment-effect estimation must be solved at a grand scale—every possible combination of interventions could be considered to find an optimal treatment policy. The diagram shows the scale of such a problem with only three distinct decisions. Blue and red people denote positive and negative outcomes, respectively. Credit: Debbie Maizels/Springer Nature

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 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 comorbid 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

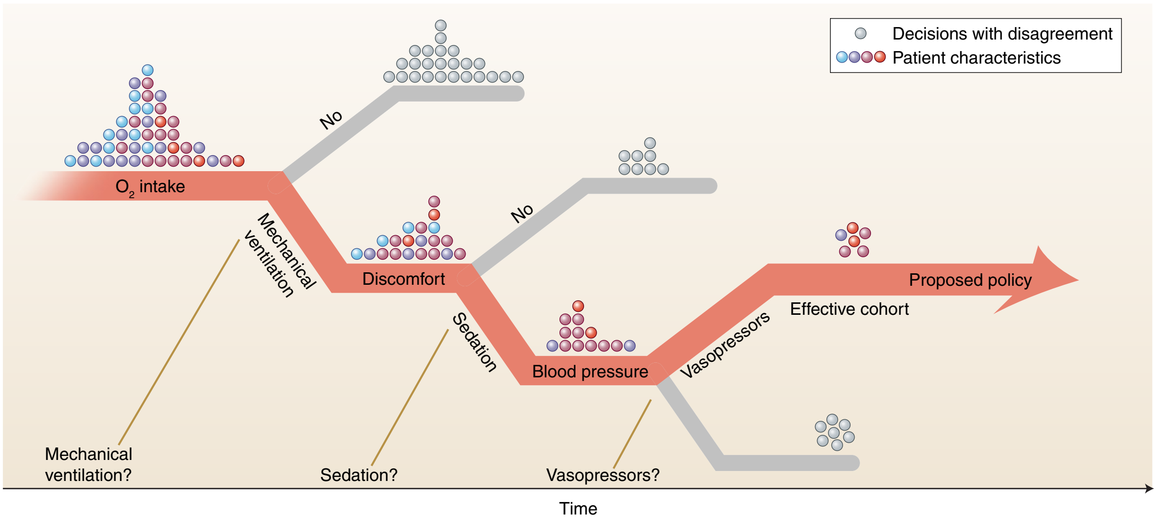

Fig. 2 | Effective sample size in off-policy evaluation. Each dot represents a single patient at each stage of treatment, and its color indicates the patient's characteristics. The more decisions that are performed in sequence, the likelier it is that a new policy disagrees with the one that was learned from. Gray decision points indicate disagreement. Use of 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. Credit: Debbie Maizels/Springer Nature

receive fluids earlier than healthier patients yet have worse outcomes, which is clearly a result of them being sicker in the first place. This difference in outcome may lead to an analysis that associates earlier fluid administration with worse outcomes if not properly adjusted for clinical context. Adjusting for confounding factors is challenging when validating the average treatment effect of a single decision[7]; this problem becomes significantly harder when decisions are made in sequence. Thus, when reading an RL study, it is important to be conscientious of possible confounding factors, even more so than for standard prediction studies, as the sequential nature of the problem could lead to confounding effects in the long as well as short term.

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 on the basis of 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 evaluation is especially prone to having small effective sample sizes (Fig. 2).

For example, we found that the effective sample size for a sepsis management policy on a cohort of 3,855 patients was only a few dozen[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 had 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, caution and due diligence must be exercised in their application and evaluation.

Omer Gottesman[1,11], Fredrik Johansson[2,11], Matthieu Komorowski[3,4], Aldo Faisal[5,6,7,8], David Sontag[2], Finale Doshi-Velez[1] and Leo Anthony Celi[3,9,10]*

[1]Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [2]Institute for Medical Engineering and Science, MIT, Cambridge, MA, USA. [3]Laboratory for Computational Physiology, Harvard-MIT Health Sciences & Technology, MIT, Cambridge, MA, USA. [4]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. [5]Department of Bioengineering, Imperial College London, London, UK. [6]Department of Computing, Imperial College London, London, UK. [7]Data Science Institute, London, UK. [8]MRC London Institute of Clinical Sciences, London, UK. [9]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [10]MIT Critical Data, Cambridge, MA, USA. [11]These authors contributed equally: Omer Gottesman, Fredrik Johansson.
**e-mail: lceli@mit.edu*

Published online: 7 January 2019
https://doi.org/10.1038/s41591-018-0310-5

References

1. Obermeyer, Z. & Emanuel, E. J. *N. Engl. J. Med.* **375**, 1216 (2016).
2. Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V. & Doshi-Velez, F. *AMIA Summits on Translational Science Proceedings* **2017**, 239 (2017).
3. Guez, A., Vincent, R. D., Avoli, M. & Pineau, J. Treatment of epilepsy via batch-mode reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* 1671–1678 (AAAI, 2008).
4. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. & Faisal, A. *Nat. Med.* **24**, 1716–1720 (2018).
5. Chakraborty, B., Moodie, E. & Erica, E. M. *Statistical Methods for Dynamic Treatment Regimes* (Springer, New York, 2013).
6. Simpson, N., Lamontagne, F. & Shankar-Hari, M. *Curr Opin Crit Care.* **23**, 561–566 (2017).
7. Johansson, F., Shalit, U. & Sontag, D. Learning representations for counterfactual inference. In *Proceedings of the 33th International Conference on Machine Learning* (ICML, 2016).
8. Precup, D., Sutton, R. S. & Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning* 759–766 (ICML, 2000).
9. Gottesman, O. et al. Evaluating Reinforcement Learning Algorithms in Observational Health Settings. Preprint at https://arxiv.org/abs/1805.12298 (2018).
10. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at https://arxiv.org/abs/1702.08608 (2017).

Competing interests

A.A.F. has received funding from Fresenius-KABI in the past.