https://doi.org/10.1038/s41591-018-0213-5

nature medicine

The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care

Matthieu Komorowski[1,2,3], Leo A. Celi[3,4], Omar Badawi[3,5,6], Anthony C. Gordon[1*] and A. Aldo Faisal[2,7,8,9*]

Sepsis is the third leading cause of death worldwide and the main cause of mortality in hospitals[1–3], but the best treatment strategy remains uncertain. In particular, evidence suggests that current practices in the administration of intravenous fluids and vasopressors are suboptimal and likely induce harm in a proportion of patients[1,4–6]. To tackle this sequential decision-making problem, we developed a reinforcement learning agent, the Artificial Intelligence (AI) Clinician, which extracted implicit knowledge from an amount of patient data that exceeds by many-fold the life-time experience of human clinicians and learned optimal treatment by analyzing a myriad of (mostly suboptimal) treatment decisions. We demonstrate that the value of the AI Clinician's selected treatment is on average reliably higher than human clinicians. In a large validation cohort independent of the training data, mortality was lowest in patients for whom clinicians' actual doses matched the AI decisions. Our model provides individualized and clinically interpretable treatment decisions for sepsis that could improve patient outcomes.

Sepsis is defined as severe infection leading to life-threatening acute organ dysfunction[7]. The management of intravenous fluids and vasopressors in sepsis is a key clinical challenge and a top research priority[1,4]. Besides general guidelines, such as the Surviving Sepsis Campaign, no tool currently exists to personalize treatment of sepsis and assist clinicians in making decisions in real-time and at the patient level[4–6]. As a consequence, clinical variability in sepsis treatment is extreme, with consistent evidence that suboptimal decisions lead to poorer outcomes[8–10].

We developed the AI Clinician, a computational model using reinforcement learning, which is able to dynamically suggest optimal treatments for adult patients with sepsis in the intensive care unit (ICU). Reinforcement learning is a category of AI tools in which a virtual agent learns from trial-and-error an optimized set of rules—a policy—that maximizes an expected return[11,12]. Similarly, a clinician's goal is to make therapeutic decisions in order to maximize a patient's probability of a good outcome[12,13]. Reinforcement learning has many desirable properties that may help medical decision-making. The intrinsic design of models using reinforcement learning can handle sparse reward signals, which makes them well-suited to overcome the complexity related to the heterogeneity of patient responses to medical interventions and the delayed indications of the efficacy of treatments[11]. Importantly, these algorithms are able to infer optimal decisions from suboptimal training examples. Reinforcement learning has been successfully applied in the past to medical problems, such as diabetes and mechanical ventilation in the ICU[14–17].

Our AI Clinician was built and validated on two large nonoverlapping ICU databases containing data routinely collected from adult patients in the United States. The Medical Information Mart for Intensive Care version III (MIMIC-III)[18] was used for model development, and the eICU Research Institute Database (eRI) for model testing. In both datasets, we included adult patients fulfilling the international consensus sepsis-3 criteria[7]. After exclusion of ineligible cases, we included 17,083 admissions (88.4% of eligible patients with sepsis) from five separate ICUs in one tertiary teaching hospital and 79,073 admissions (73.6% of eligible patients with sepsis) from 128 different hospitals from MIMIC-III and eRI, respectively (Supplementary Fig. 1). Patient demographics and clinical characteristics are shown in Table 1 and Supplementary Table 1.

In both datasets, we extracted a set of 48 variables, including demographics, Elixhauser premorbid status[19], vital signs, laboratory values, fluids and vasopressors received (Supplementary Table 2). Patients' data were coded as multidimensional discrete time series with 4-h time steps, and for each patient, we included up to 72 h of measurements taken around the estimated time of onset of sepsis. The total volume of intravenous fluids and maximum dose of vasopressors administered over each 4-h period defined the medical treatments of interest. The model aims at optimizing patient mortality, so a reward was associated to survival and a penalty to death.

A Markov decision process (MDP) was used to model the patient environment and trajectories[20,21]. The various elements of the model were defined using patient data time series from the training set (a random sample of 80% of MIMIC-III; Fig. 1). We deployed the AI Clinician to solve the MDP and predict outcomes of treatment strategies. First, we evaluated the actual treatments of clinicians by analyzing all the prescriptions and computing the average return of each treatment option, which can take values from −100 to + 100 in our model. Then, the MDP was solved using policy iteration, which identified the treatments that maximized return, that is, the expected 90-d survival of patients in the MIMIC-III cohort[11]. The resultant policy is referred to hereafter as the 'AI policy'.

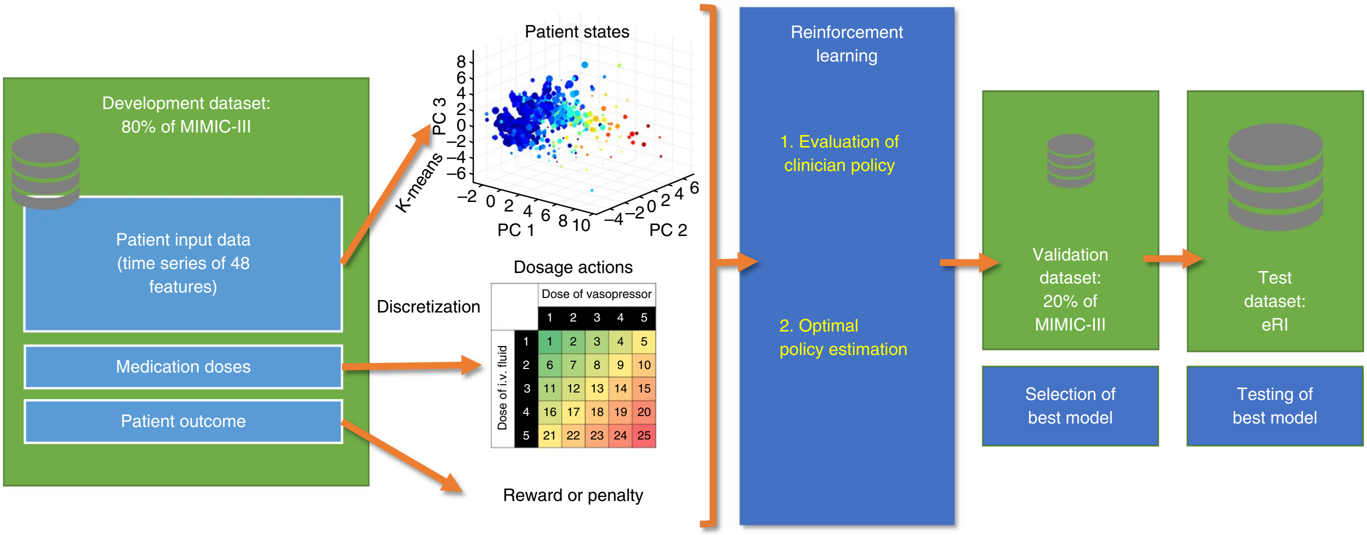
Evaluating the performance of this new AI policy using the trajectories of patients generated by another policy (the clinicians' policy)

[1]Department of Surgery and Cancer, Imperial College London, London, UK. [2]Department of Bioengineering, Imperial College London, London, UK. [3]Laboratory of Computational Physiology, Harvard–MIT Division of Health Sciences & Technology, Cambridge, MA, USA. [4]Beth Israel Deaconess Medical Center, Boston, MA, USA. [5]Department of eICU Research and Development, Philips Healthcare, Baltimore, MD, USA. [6]Department of Pharmacy Practice and Science, University of Maryland, School of Pharmacy, Baltimore, MD, USA. [7]Department of Computer Science, Imperial College London, London, UK. [8]Medical Research Council London Institute of Medical Sciences, London, UK. [9]Behaviour Analytics Lab, Data Science Institute, London, UK.
*e-mail: anthony.gordon@imperial.ac.uk; a.faisal@imperial.ac.uk

Table 1 | Description of the datasets

	MIMIC-III	eRI
Unique ICUs (*n*)	5	128
Unique ICU admissions (*n*)	17,083	79,073
Characteristics of hospitals, per number of ICU admissions	Teaching tertiary hospital	Nonteaching: 37,146 (47.0%) Teaching: 29,388 (37.2%) Unknown: 12,539 (15.9%)
Age, years (mean (s.d.))	64.4 (16.9)	65.0 (16.7)
Male gender (*n* (%))	9,604 (56.2%)	40,949 (51.8%)
Premorbid status (*n* (%))		
Hypertension	9,384 (54.9%)	43,365 (54.8%)
Diabetes	4,902 (28.7%)	25,290 (32.0%)
CHF	5,206 (30.5%)	15,023 (19.0%)
Cancer	1,803 (10.5%)	11,807 (14.9%)
COPD or RLD	4,248 (28.7%)	18,406 (23.3%)
CKD	3,087(18.1%)	14,553 (18.4%)
Primary ICD-9 diagnosis (*n* (%))		
Sepsis, including pneumonia	5,824 (34.1%)	41,396 (52.3%)
Cardiovascular	5,270 (30.8%)	11,221 (14.2%)
Respiratory	1,798 (10.5%)	9,127 (11.5%)
Neurological	1,590 (9.3%)	7,127 (9.0%)
Renal	429 (2.5%)	1,454 (1.8%)
Others	2,172 (12.7%)	8,747 (11.1%)
Initial OASIS (mean (s.d.))	33.5 (8.8)	34.8 (12.4)
Initial SOFA (mean (s.d.))	7.2 (3.2)	6.4 (3.5)
Procedures during the 72 h of data collection:		
Mechanical ventilation (*n* (%))	9,362 (54.8%)	39,115 (49.5%)
Vasopressors (*n* (%))	6,023 (35.3%)	23,877 (30.2%)
Renal replacement therapy (*n* (%))	1,488 (8.7%)	6,071 (7.7%)
Length of stay, days (median, (IQR))	3.1 (1.8–7)	2.9 (1.7–5.6)
ICU mortality	7.4%	9.8%
Hospital mortality	11.3%	16.4%
90-d mortality	18.9%	Not available

CHF, congestive heart failure; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; ICD-9, International Classification of Diseases version 9; IQR, interquartile range; OASIS, Oxford Acute Severity of Illness Score; RLD, restrictive lung disease; SOFA, sequential organ failure assessment.



	1	2	3	4	5
1	1	2	3	4	5
2	6	7	8	9	10
3	11	12	13	14	15
4	16	17	18	19	20
5	21	22	23	24	25

Fig. 1 | Data flow of the AI Clinician. Eighty percent of the MIMIC-III dataset was used to define the elements of the MDP. Time series of patient data were clustered into finite states. The dose of intravenous (i.v.) fluids and vasopressors were discretized into 25 possible actions. Patient survival at 90 d after ICU admission defined reward. Reinforcement learning was used to estimate optimal treatment strategies—the AI policy. The remaining 20% of MIMIC-III data was used to identify the best model among 500 candidates, which was then tested on an independent dataset from the eRI database.

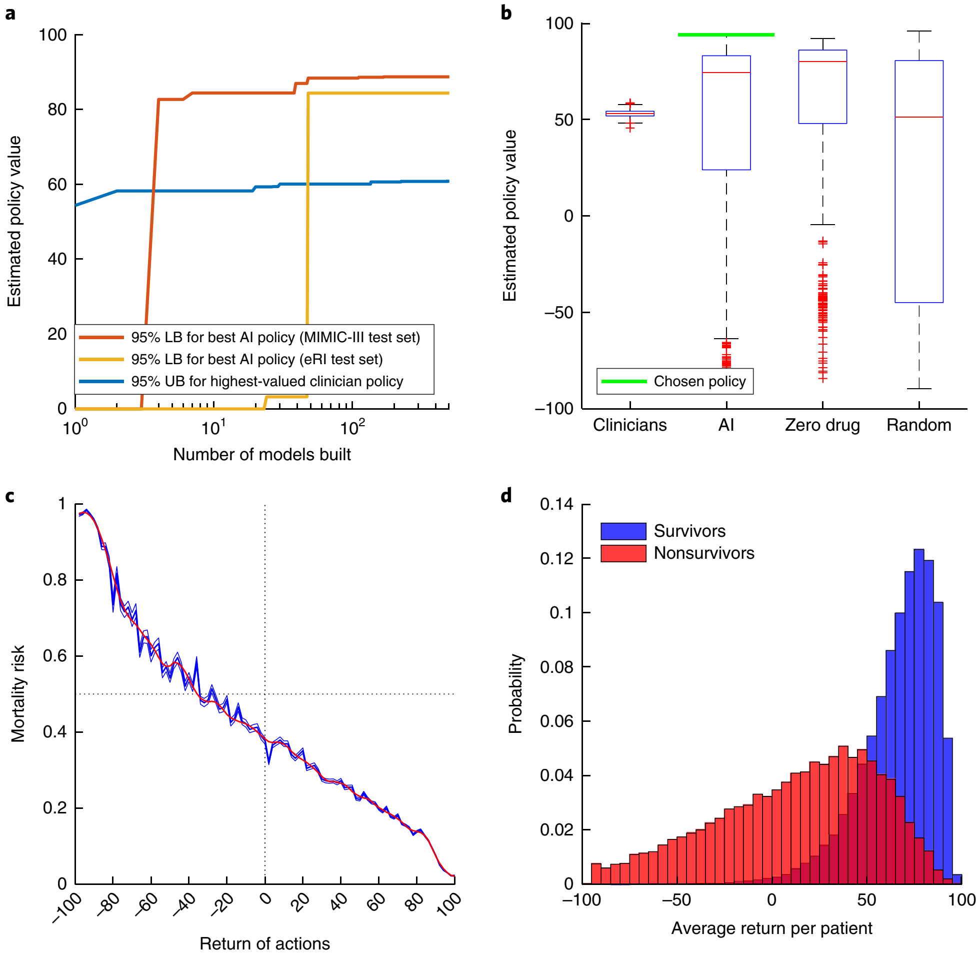


Fig. 2 | Selection of the best AI policy and model calibration. **a**, Evolution of the 95% lower bound (LB) of the best AI policy and of the 95% upper bound (UB) of highest-valued clinician policy during the building of 500 models. After only a few models, a higher value for the AI policy than the clinician treatment, within the accepted risk, is guaranteed. $n = 13{,}666$ patients in the MIMIC-III development dataset, $n = 3{,}417$ in the MIMIC-III test set and $n = 79{,}073$ in the eRI set. **b**, Distribution of the estimated value of the clinicians' actual treatments, the AI policy, a random policy and a zero-drug policy across the 500 models in the MIMIC-III test set ($n = 500$ models in each boxplot). The chosen AI policy maximizes the 95% confidence lower bound. On each boxplot, the central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range. Points beyond the whiskers are considered outliers and are plotted individually using the + symbol. **c**, The relationship between the return of clinicians' treatments and patient 90-d mortality in the MIMIC-III training set ($n = 13{,}666$ patients). Return of actions were sorted into 100 bins, and the mean observed mortality (blue line for raw, red line for smoothed) was computed in each of these bins. The shaded blue area represents the s.e.m. Treatments with a low return were associated with a high risk of mortality, whereas treatments with a high return led to better survival rates. **d**, Average return in survivors ($n = 11{,}031$) and nonsurvivors ($n = 2{,}635$) in the MIMIC-III training set. **c** and **d** were generated by bootstrapping in the training data with 2,000 resamplings.

is termed off-policy evaluation[22–24]. It was crucial to obtain reliable estimates of the performance of this new policy without deploying it, as executing a bad policy would be dangerous for patients[22,23]. Therefore, we implemented a type of high-confidence off-policy evaluation (HCOPE) method (weighted importance sampling (WIS)), and we used bootstrapping to estimate the true distribution of the policy value in the MIMIC-III 20% validation set (Fig. 2b and Supplementary Fig. 1)[23,24]. We built 500 different models using 500 different clustering solutions of the training data, and the selected final model maximized the 95% confidence lower bound of the AI policy[23]. Fig. 2a shows that this bound consistently exceeded the 95% confidence upper bound of the clinicians' policy, provided that enough models were built. This model selection method maximizes the theoretical statistical safety of the new AI policy. The chosen AI policy was then tested on the independent eRI dataset.

Good model calibration was confirmed by plotting the relationship between the return of the clinicians' policy and patients' 90-day mortality (Fig. 2c). In Fig. 2d, we show the average return measured in survivors and nonsurvivors.

Fig. 3a shows the distribution of the estimated value of the clinicians' policy and the AI policy in the selected final model tested on the eRI cohort. Using bootstrapping with 2,000 resamplings, the median value of clinicians' policy and the AI policy were estimated at 56.9 (interquartile range, 54.7–58.8) and 84.5 (interquartile range, 84.3–87.7), respectively. Fig. 3b,c shows the distribution of treatment doses according to clinicians' and AI policies. On average, the AI Clinician recommended lower doses of intravenous fluids and higher doses of vasopressors than the clinicians' actual treatments. The proportion of time the eRI patients received vasopressors was only 17%, but this would have been 30% if the AI Clinician's recommendation was followed.

We further validated the model by analyzing patient mortality when the dose actually administered corresponded to or differed from the dose suggested by the AI Clinician. Fifty-eight percent

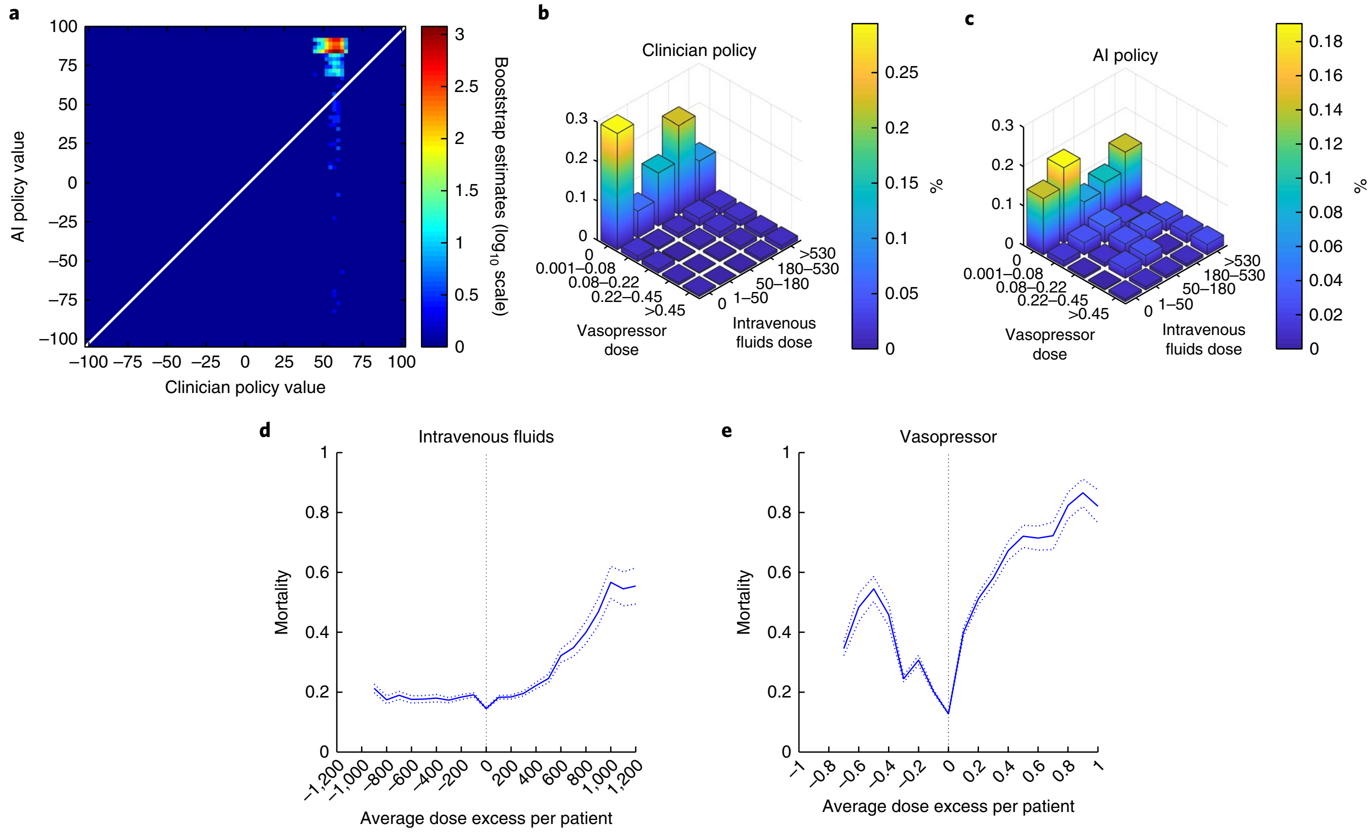


Fig. 3 | Comparison of clinician and AI policies in eRI and average dose excess received per patient of both drugs in eRI with corresponding mortality. a, Distribution of the estimated value of the clinician and AI policies in the selected model, built by bootstrapping with 2,000 resamplings. **b,c**, Visualization of the clinician and AI policies. All actions were aggregated over all time steps for the five dose bins of both medications. On average, patients were administered more intravenous fluid (**b**) and less vasopressor (**c**) medications than recommended by the AI policy. Vasopressor dosage is shown in μg/kg/min of norepinephrine equivalent, and intravenous fluid dosage is shown in mL/4 h. **d**, The dose excess, referring to the difference between the given and suggested dose averaged over all time points per patient, for intravenous fluids (left) and vasopressor (right). The figure was generated by bootstrapping with 2,000 resamplings. The shaded area represents the s.e.m. In both plots, the smallest dose difference was associated with the best survival rates (vertical dotted line). The further away the dose received was from the suggested dose, the worse the outcome.

of the time, the patients received a dose of vasopressor very close to the suggested dose, within 0.02 μg/kg body weight/min (μg/kg/min) or 10% (whichever was smaller). For fluids, patients received the suggested dose approximately 36% of the time, within 10 mL/hour or 10%. These patients, who received doses similar to the doses recommended by the AI Clinician, had the lowest mortality. When the actual dose given was different from the suggested dose, clinicians gave more or less fluids in similar proportions and less vasopressor 75% of the time. Administering more or less of either treatment than the AI policy was associated with increasing mortality rates in a dose-dependent fashion. Fig. 3d,e depicts this association, with the dose gap averaged at the patient level. The median dose deficit in patients who received too little vasopressor was 0.13 μg/kg/min (interquartile range, 0.04–0.27 μg/kg/min).

Using a random forest classification model, we gained some insight into the model representations and interpretability by estimating the relative importance of the model parameters for predicting the administration of both medications (Supplementary Fig. 2). This confirmed that the decisions suggested by the AI Clinician were clinically interpretable and relied primarily on sensible clinical and biological parameters.

Here we demonstrate how reinforcement learning could be applied to solve a complex medical problem and suggest individualized and clinically interpretable treatment strategies for sepsis. In an independent cohort, the patients who received the treatments suggested by the AI Clinician had the lowest mortality rate.

When clinicians' actual treatments varied from the AI Clinician's suggested policy, this was most commonly administration of too little vasopressor. Early use of low-dose vasopressor has been suggested to play a role in sepsis;[4,5,8,9] this may avoid administration of an excessive amounts of fluids, which has been linked with a poorer outcome[1,4,5,25]. Our results support this strategy but importantly allow the treatment to be individualized for each patient.

We envision that this system would be used in real-time, with patient data obtained from different streams being fed into electronic health record software fitted with our algorithm, which would suggest a course of action. Physicians will always need to make subjective clinical judgments about treatment strategies, but computational models can provide additional insight about optimal decisions, avoiding targeting short-term resuscitation goals and instead following trajectories toward longer-term survival[26–28]. The reinforcement learning approach that we have developed is agnostic to data used and could in principle be applied to any data-rich clinical environment and many medical interventions. In the future, it is likely that as '-omic' technologies develop, this additional information will be added to the AI Clinician to improve state definition and guide more therapies in selected patient groups.

However, there are limitations to our study. Although the datasets we used are large and comprise routinely collected clinical data, some sites and patients had to be excluded owing to poor-quality data recording or missing data. Because of differences between the two datasets, slightly different implementations of the sepsis-3 criteria were used, and hospital mortality was used in eRI instead of 90-d mortality. Finally, some laboratory values would not have been immediately available to clinicians to inform decision-making but were available to the AI Clinician.

This work will clearly require prospective evaluation using real-time data and decision-making in clinical trials and also testing in different healthcare settings, but a reduction in mortality from sepsis by only a small percentage would represent several tens of thousands of lives saved annually worldwide[3]. In the last 10–15 years, attempts to develop new treatments to reduce sepsis mortality have uniformly been unsuccessful[29,30]. The use of computer decision support systems to better guide treatments and improve outcomes is therefore a much needed approach.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0213-5.

Received: 23 March 2018; Accepted: 13 August 2018;
Published online: 22 October 2018

References

1. Gotts, J. E. & Matthay, M. A. Sepsis: pathophysiology and clinical management. *Br. Med. J.* **353**, i1585 (2016).
2. Torio, C. M. & Andrews, R. M. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011: Statistical Brief #160. in *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs* (Agency for Health Care Research and Quality, Rockville, MD, USA, 2013).
3. Liu, V. et al. Hospital deaths in patients with sepsis from 2 independent cohorts. *J. Am. Med. Assoc.* **312**, 90–92 (2014).
4. Byrne, L. & Van Haren, F. Fluid resuscitation in human sepsis: time to rewrite history? *Ann. Intensive Care* **7**, 4 (2017).
5. Marik, P. E. The demise of early goal-directed therapy for severe sepsis and septic shock. *Acta Anaesthesiol. Scand.* **59**, 561–567 (2015).
6. Marik, P. & Bellomo, R. A rational approach to fluid therapy in sepsis. *Br. J. Anaesth.* **116**, 339–349 (2016).
7. Singer, M. et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *J. Am. Med. Assoc.* **315**, 801–810 (2016).
8. Waechter, J. et al. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. *Crit. Care Med.* **42**, 2158–2168 (2014).
9. Bai, X. et al. Early versus delayed administration of norepinephrine in patients with septic shock. *Crit. Care.* **18**, 532 (2014).
10. Marik, P. E., Linde-Zwirble, W. T., Bittner, E. A., Sahatjian, J. & Hansell, D. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. *Intensive Care Med.* **43**, 625–632 (2017).
11. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. 1st edn (MIT Press, Cambridge, MA, USA, 1998).
12. Bennett, C. C. & Hauser, K. Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. *Artif. Intell. Med.* **57**, 9–19 (2013).
13. Schaefer, A. J., Bailey, M. D., Shechter, S. M. & Roberts, M. S. Modeling Medical Treatment Using Markov Decision Processes. in *Operations Research and Health Care* (eds. Brandeau, M. L., Sainfort, F. & Pierskalla, W. P.) 593–612 (Springer, Boston, 2005).
14. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
15. Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M. & Engelhardt, B. E. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. Preprint at https://arxiv.org/abs/1704.06300 (2017).
16. Bothe, M. K. et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert. Rev. Med. Devices.* **10**, 661–673 (2013).
17. Lowery, C. & Faisal, A. A. Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control. in *International IEEE/EMBS Conference on Neural Engineering* 1414–1417 (IEEE, San Diego, CA, USA, 2013).
18. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
19. Elixhauser, A., Steiner, C., Harris, D. R. & Coffey, R. M. Comorbidity measures for use with administrative data. *Med. Care* **36**, 8–27 (1998).
20. Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. (Wiley-Interscience, Hoboken, NJ, USA, 2014).
21. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. 2nd edn,(MIT Press, Cambridge, MA, USA, 2018).
22. Thomas, P. S., Theocharous, G. & Ghavamzadeh, M. High-Confidence Off-Policy Evaluation. in *Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, CA, USA, 2015).
23. Hanna, J. P., Stone, P. & Niekum, S. Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation. Preprint at https://arxiv.org/abs/1606.06126 (2016).
24. Thomas, P. S., Theocharous, G. & Ghavamzadeh, M. High confidence policy improvement. in *Proceedings of the 32nd International Conference on Machine Learning* 2380–2388 (PMLR, Lille, France, 2015).
25. Acheampong, A. & Vincent, J.-L. A positive fluid balance is an independent prognostic factor in patients with sepsis. *Crit. Care.* **19**, 251 (2015).
26. Johnson, A. E. W. et al. Machine learning and decision support in critical care. *Proc. IEEE Inst. Electr. Electron Eng.* **104**, 444–466 (2016).
27. Vincent, J.-L. The future of critical care medicine: integration and personalization. *Crit. Care Med.* **44**, 386–389 (2016).
28. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507–2509 (2017).
29. Gordon, A. C. et al. levosimendan for the prevention of acute organ dysfunction in sepsis. *N. Engl. J. Med.* **375**, 1638–1648 (2016).
30. Ranieri, V. M. et al. Drotrecogin alfa (activated) in adults with septic shock. *N. Engl. J. Med.* **366**, 2055–2064 (2012).

Acknowledgements

We are grateful to F. Doshi-Velez and O. Gottesman for their assistance with the methodology. We are grateful for support from the National Institute of Health Research (NIHR) Comprehensive Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London. We are thankful to the Laboratory of Computational Physiology at the Massachusetts Institute of Technology and the eICU Research Institute for providing the data used in this research. M.K. and this project are funded by the Engineering and Physical Sciences Research Council and an Imperial College President's PhD Scholarship. A.C.G. is funded by an NIHR Research Professorship award (RP-2015-06-018). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions

M.K., A.C.G. and A.A.F. conceived the overall study. M.K. and A.A.F. designed and conducted the experiments and analyzed the data. L.A.C. and O.B. contributed to the experimental design and analyses. O.B. provided key input in extracting and processing data from the eRI. All authors contributed to the interpretation of the results and M.K. drafted the manuscript, which was reviewed, revised and approved by all authors.

Competing interests

The authors declare competing interests: A.C.G. reports that outside of this work he has received speaker fees from Orion Corporation Orion Pharma and Amomed Pharma. He has consulted for Ferring Pharmaceuticals, Tenax Therapeutics, Baxter Healthcare, Bristol-Myers Squibb and GSK, and received grant support from Orion Corporation Orion Pharma, Tenax Therapeutics and HCA International with funds paid to his institution. L.A.C. receives funding from Philips Healthcare. O.B. is an employee of Philips Healthcare. A.A.F. has received funding from Fresenius-KABI. M.K. does not have competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0213-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.C.G. or A.A.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study design and databases. We built and then validated a computational clinical-decision support model based on the retrospective analysis of two nonoverlapping intensive care databases containing data collected from adult patients. The databases were:

(i) The MIMIC-III, an open-access, anonymized database of 61,532 admissions from 2001–2012 in six ICUs at a Boston teaching hospital[18].
(ii) The Philips eRI, containing more than 3.3 million admissions from 2003–2016 in 459 ICUs across the United States.

MIMIC-III was used for model development, and eRI for model validation. Both databases contain high-resolution patient data including demographics, vital signs time series, laboratory tests, illness severity scores, medications and procedures, fluid intake and outputs, clinician notes, and diagnostic coding.

Patient cohorts. In both datasets, MIMIC-III and eRI, we included adult patients fulfilling the sepsis-3 criteria[7]. Sepsis was defined as a suspected infection (prescription of antibiotics and sampling of bodily fluids for microbiological culture) combined with evidence of organ dysfunction, defined by a SOFA score ≥2 (refs [7,31]). We adhered to the original temporal criteria for diagnosis of sepsis: when the antibiotic was given first, the microbiological sample must have been collected within 24 h; when the microbiological sampling occurred first, the antibiotic must have been administered within 72 h[31]. The earlier event defined the onset of sepsis. In line with previous research, we assumed a baseline SOFA of zero for all patients[31,32].

Exclusion criteria.

- In both databases:
  - Age <18 years old at the time of ICU admission
  - Mortality not documented
  - Withdrawal of treatment (see below)
- In MIMIC-III:
  - Intravenous fluid intake not documented
- In eRI:
  - ICU readmissions, because of the potential risk in this database of mixing up data from subsequent ICU admissions.
  - Patient admitted in an ICU with insufficient data collection (see below).

We excluded patients whose treatment was withdrawn because in this case clinical decisions are no longer made aiming to optimize survival, which would have led to spurious actions in the AI policy. Withdrawal of treatment often involves patients with high severity of illness and who are on high doses of vasopressors, in which treatment is withdrawn as it is considered futile. Therefore, we defined withdrawal as patients who died within 24 h of the end of the data collection period and received vasopressors at any point and whose vasopressors were stopped at the end of the data collection.

In the eRI, the data was recorded heterogeneously across ICUs. To avoid any systematic bias in our analysis (for example, when no medication appears in the database, where in reality it was actually administered to the patient), we excluded hospitals for the years in which the availability of data was not sufficient, as data recording practices could vary over time. We defined two indicators of data availability for vasopressors and intravenous fluids, averaged per day, per patient, per hospital and per year. Given that our analysis resolution was 4 h, we expected at least six records per day, even if the dose was constant. Hospital-years with less than six daily records on average were excluded. In total, 331 ICUs out of 459 were excluded with the combined data-quality-selection approach. For comparison, the data quality in MIMIC-III was high, with a weighted daily average over the five ICUs of 20.4 intravenous fluids records and 31.1 vasopressor records.

Data extraction and preprocessing. In MIMIC-III, data were included from up to 24 h preceding until 48 h following the estimated onset of sepsis, in order to capture the early phase of its management, including initial resuscitation. The outcome was 90-day mortality. Owing to the size of the eRI database (over 2.4 terabytes), a simplified data-extraction process had to be employed. Therefore, we identified all adult patients who had sepsis during the first 36 h after admission and extracted the data obtained from these patients over the first 72 h after admission. Survival at 90 d was not available in eRI, so hospital mortality defined the outcome of interest in this cohort.

From both datasets, we extracted a set of 48 variables, including demographics, Elixhauser premorbid status[19], vital signs, laboratory values, fluids and vasopressors received and fluid balance (Supplementary Table 2). Patients' data were coded as multidimensional discrete time series with 4-h time steps. Data variables with multiple measurements within a 4-h time step were averaged (for example, heart rate) or summed (for example, urine output) as appropriate.

All features were checked for outliers and errors using a frequency histogram method and univariate statistical approaches (Tukey's method). Errors were corrected when possible (for example, conversion of temperature from Fahrenheit to Celsius degrees). Parameters were capped to clinically plausible values.

To address the problem of missing or irregularly sampled data, we used a time-limited parameter-specific sample-and-hold approach in both datasets, a common practice with health time series data that intuitively matches the cognitive process of clinicians[33]. The remaining missing data were interpolated in MIMIC-III using multivariable nearest-neighbor imputation as the clustering algorithm did not accept missing values[34]. We did not interpolate the remaining missing data in eRI as it was not required for model validation.

Building the computational model. The true patient physiological state is only partially represented by the data available, and therefore the disease process could be formulated as a partially observable MDP. A MDP was used to approximate patient trajectory and to model the decision-making process[12,20,21]. The MDP is defined by the tuple $\{S, A, T, R, \gamma\}$, with:

- S, a finite set of states (in our model, the health states of patients).
- A, the finite set of actions available from state s (in our model, the dose prescribed of intravenous fluids and vasopressors converted into discrete decisions).
- $T(s',s,a)$, the transition matrix, containing the probability that action a in state s at time t will lead to state s' at time $t+1$, which describes the dynamics of the system.
- $R(s')$, the immediate reward received for transitioning to state s'. Transitions to desirable states yield a positive reward, and reaching undesirable states generates a penalty.
- γ, the discount factor, which allows modelling of the fact that a future reward is worth less than an immediate reward.

A sample of 80% of the MIMIC-III cohort was used for model training, and the remaining 20% was used for model validation. The state space was defined by clustering all patient time series from the MIMIC-III development set.
A good cluster hierarchy is one in which individuals that are in the same cluster are similar with respect to their observable properties. Specifically, the state space was constructed by k-means++ clustering of the patients' data, resulting in 750 discrete mutually exclusive patient states[35]. We used Bayesian and Akaike information criteria to determine the optimal number of clusters (Supplementary Fig. 3e)[36]. We chose a high value of k to ensure a highly granular model while avoiding using too large a state space (>1,000), which would have led to very sparsely populated states (Supplementary Fig. 3a). Two absorbing states were added to the state space, corresponding to death and discharge of the patient. To further assess the validity of our state aggregation, we used the distribution of International Classification of Diseases codes in the states and demonstrated that past medical history and diagnoses are encapsulated to some extent within our chosen state definition (Supplementary Fig. 3b).

Prior to clustering and to account for unequal means and variances in data, normally distributed data was standardized, log-normal distributed variables were log-transformed before standardization, and binary data was centered on zero. The normality of each variable was tested with visual methods: quantile-quantile plots and frequency histograms.

The management of ICU patients with sepsis is extremely complex and includes several principles such as rapid control of the source of infection, correction of hypovolaemia, and management of secondary organ failures. Including all these potential interventions as actions in the MDP would have required a much larger dataset. A key challenge is arguably the management of intravenous fluids and vasopressors. Consequently, we focused on medical decisions regarding total volume of intravenous fluids and maximum dose of vasopressors administered over each 4-h period. Intravenous fluids included boluses and background infusions of crystalloids, colloids and blood products, normalized by tonicity as previously described[8]. The vasopressors included norepinephrine, epinephrine, vasopressin, dopamine and phenylephrine and were converted when necessary to norepinephrine-equivalent using previously published dose correspondence[37]. To define the action space, the dose of each treatment was represented as one of five possible choices, choice 1 being 'no drug given' and the remaining non-null doses divided into four quartiles (Supplementary Table 3). The combination of the two treatments produced 25 possible discrete actions. We expressed the suggested dose as the median of each dose bin matching a suggested action.

The sequences of successive states and actions are referred to as patients' trajectories. In our models, we used either hospital mortality or 90-d mortality as the sole defining factor for the system-defined penalty and reward. When a patient survived, a positive reward was released at the end of each patient's trajectory (a 'reward' of +100); a negative reward (a 'penalty' of −100) was issued if the patient died.

We estimated the transition matrix $T(s',s,a)$ by counting how many times each transition was observed in the MIMIC-III training dataset and converting the transition counts to a stochastic matrix[32]. In high-risk environments (where executing a bad policy could cause harm) limiting the action space to known options is a sensible choice to increase the safety of the model. We restricted the set of actions to choose from to frequently observed actions taken by clinicians and excluded transitions seen fewer than five times. As such, the resulting AI policy suggests the best possible treatment among all the options chosen (relatively frequently) by clinicians.

Markov models rely on the Markov property, which is that the transitions (given state and action) are memoryless. The probability to leave a state in a Markov chain remains constant, no matter how long the agent has been in the state. Thus, the probability to remain in a state follows an exponential decay[38]. We measured the empirical state persistence probability for each state and found a high goodness-of-fit between the data and exponential decay distributions for virtually all states (Supplementary Fig. 3).

The discount factor γ defines the horizon of the reinforcement learning agent, which is how much importance is given to future rewards compared to the reward in the current state. It can take values between 0 and 1 (ref. [21]). We chose a γ value of 0.99, which means that we put nearly as much importance on late deaths as opposed to early deaths.

In reinforcement learning, a policy π corresponds to a set of rules dictating which action is taken while in a particular state[21]. Each MDP determines a state-action value function Q^π, that reflects the expected sum of discounted rewards for choosing an action while in a particular state, and following a policy π thereafter[21]. In our model, Q^π summarizes the effect of the treatment decisions on the patient's mortality risk, with beneficial decisions having positive Q^π values and harmful decisions negative Q^π values[12,13].

Evaluation of clinicians' actions. We performed an evaluation of the actual actions (the policy) of clinicians using temporal difference learning (TD-learning) of the *Q* function by observing all the prescriptions of intravenous fluids and vasopressors in existing records (offline sampling) and computing the average value of each treatment option at the state level[21]. The advantage of TD-learning over policy iteration is that it does not require knowledge of the MDP (model-free) and makes it possible to learn simply from sample trajectories[21]. It was computed iteratively from actual patient episodes of successive state-action pairs, with resampling, using the following Q update formula:

$$Q^\pi(s,a) \leftarrow Q^\pi(s,a) + \alpha \cdot (r + \gamma \cdot Q^\pi(s',a') - Q^\pi(s,a)) \quad (1)$$

With $Q^\pi(s,a)$ the current {state, action} tuple considered, $Q^\pi(s',a')$ the next {state, action} tuple, α the learning rate and *r* the immediate reward.

We stopped the evaluation after processing 500,000 patient trajectories with resampling, which is when the value of the estimated policy reached an asymptote.

Estimation of the AI policy. We learned a theoretical optimal policy (which we call the 'AI' policy) for the MDP using in-place policy iteration, which identified the decisions that maximize the long-term sum of rewards, hence the expected survival of patients[21]. Policy iteration started with a random policy that was iteratively evaluated and then improved until converging to an optimal solution. After convergence, the AI policy π^*corresponded to the actions with the highest state-action value in each state:

$$\pi^*(s) \leftarrow \underset{a}{\operatorname{argmax}}\, Q^{\pi^*}(s,a) \ \ \forall s \quad (2)$$

The value *V* of a policy π was computed using the Bellman equation for V^π and represented the expected return when starting in *s* and following π thereafter:

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} T(s',s,a)\,[R(s') + \gamma V^\pi(s')] \quad (3)$$

Because 90-d mortality was not available in the eRI, hospital mortality was used as the outcome of interest. We verified first that the model performed well in the MIMIC-III database, when the model was trained using hospital mortality (Supplementary Fig. 4) and 90-d mortality (data not shown). This sanity check supported the extension of the framework into the eRI data.

Model evaluation. Our objective is to evaluate the value of a newly learnt AI policy using trajectories of patients generated by another policy (the clinicians')[21–23]. This is termed off-policy evaluation (OPE). Using direct, model-based estimates of the policy value are known to reduce variance at the cost of adding bias to the estimate[22,39]. Therefore, we implemented a type of HCOPE method, WIS, and used bootstrapping to estimate the true distribution of the policy value in the test sets[22,23,40]. WIS may be a biased although consistent policy estimator, so the bootstrap confidence interval may also be biased, even though the literature suggests that consistency is a more desirable property than unbiasedness[22,39,41]. It is accepted that bootstrapping produces accurate confidence intervals with less data than exact HCOPE methods and is safe enough in high-risk applications, such as healthcare[22,23]. Of note, the use of bootstrap confidence intervals around WIS estimates has not been previously described in biomedical research, but the approach is suggested in reinforcement learning research[22,23].

We define π_0 as the behavior policy (the clinicians'), from which actual patient data was generated, and π_1 as the evaluation, or AI policy. In OPE tasks, importance sampling is a simple way to correct for the discrepancy between π_0 and π_1 (ref. [42]). Weighting the estimate allows reducing its variance[39]. Using importance sampling (IS) methods with a deterministic evaluation policy is problematic, as only a few {s,a} pairs and short sequences can be used for policy evaluation. Indeed, the IS weights become zero as soon as the two policies diverge. We softened π_1, so it now recommends taking the suggested action 99% of the time and any of the other actions a total of 1% of the time. This allows assessment of the entirety of the patient trajectories. Our goal was to estimate the value of π_1 from data trajectories. We defined $\rho_t := \pi_1(a_t\,|s_t)/\pi_0(a_t\,|s_t)$ as the per-step importance ratio, $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$ as the cumulative importance ratio up to step t, $w_t = \sum_{i=1}^{|D|} \rho_{1:t}^{(i)}/|D|$ as the average cumulative importance ratio at horizon *t* in dataset *D* and $|D|$ as the number of trajectories in *D*[21,39]. The trajectory-wise WIS estimator is given by:

$$V_{WIS} = \frac{\rho 1:H}{w_H}\left(\sum_{t=1}^{H} \gamma^{t-1} r_t\right) \quad (4)$$

Then, the WIS estimator is the average estimate over all trajectories, namely:

$$WIS = \frac{1}{|D|} \sum_{i=1}^{|D|} V_{WIS}^{(i)} \quad (5)$$

Where $V_{WIS}^{(i)}$ is WIS applied to the *i*-th trajectory.

We built 500 different models from various selections of a random 80% of the MIMIC-III data and evaluated the AI policies with WIS on the remaining 20% of the data. State membership for test set data points was determined according to whichever training set cluster centroid they fell closest to. Once the state membership was known, we knew what the suggested action and its corresponding recommended dose of medications were. In each model, we also estimated the value of a random policy and a zero-drug policy for comparison (Fig. 2b). As recommended, the selected final model maximizes the 95% confidence lower bound of the AI policy among the 500 candidate models[22]. We demonstrate that this bound consistently exceeded the 95% confidence upper bound of the clinicians' policy, provided that enough models were built (Fig. 2a). Then, we tested the selected policy in the eRI (Fig. 3a).

We also tested the influence of the variability of the behavior policy by measuring the WIS estimator using 500 different behavior policies (generated by 500 different clustering of the training data) but a fixed evaluation policy. The 95% lower bound of the AI policy exceeded the 95% upper bound of the clinicians' policy 66.4% of the time. Considering that we selected the AI policy maximizing the WIS estimator, the models for which the variability in the behavior policy led to a low WIS estimator were discarded owing to design.

Because laboratory results are recorded in the data at the time of sampling, a fraction of the laboratory values would not have been immediately available to clinicians to inform decision-making but were available to the AI Clinician. We tested the effect of this potential bias by artificially shifting all the 'slow' laboratory tests (white blood cell count, platelet count, clotting, renal and liver function tests, etc.) in the eRI cohort 4 h into the future. This manipulation did not significantly alter the WIS estimator: 85.1 (interquartile range, 85.1–86.0; *t*-test $P > 0.05$).

Similarly to previous work, we also measured the performance of the AI policy using direct indicators and analyzed patient outcomes as a function of the gap between clinicians and AI policies[15]. Here, we analyzed patient mortality in the test sets for which the dose actually administered corresponded to or differed from the dose suggested by the AI policy (Fig. 3d,e and Supplementary Fig. 4). We used bootstrapping to generate confidence bounds.

Human subject data. The institutional review board (IRB) of the Massachusetts Institute of Technology (no. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) approved the use of MIMIC-III for research. The use of the eRI database was approved by the eICU research committee and exempt from IRB approval as the database security schema and the reidentification risk were certified as meeting safe harbor standards by Privacert (Cambridge, MA) (45 Code of Federal Regulations 164.514(b)(1) and Health Insurance Portability and Accountability Act Certification no. 1031219-2). Because this study was a secondary analysis of fully anonymized data, individual patient consent was not required.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

MIMIC-III is openly available. Access to the eRI data is restricted to the Philips eICU Research Institute. The eICU Collaborative Research Database contains a sample of over 200,000 patient stays from the eRI database that is freely available. The databases were queried in pgAdmin 4 v 1.3, and computations were implemented in Matlab R2017a (MathWorks, Inc.). Access to the computer code used in this research is available by request to the corresponding authors. To facilitate the reproduction of our results, we provide the list of anonymous patient identifiers for both databases in Supplementary Data 1 and 2.

References

31. Seymour, C. W. et al. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *J. Am. Med. Assoc.* **315**, 762–774 (2016).

32. Raith, E. P. et al. Prognostic accuracy of the SOFA Score, SIRS Criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *J. Am. Med. Assoc.* **317**, 290–300 (2017).
33. Hug, C. W. *Detecting hazardous intensive care patient episodes using real-time mortality models*. PhD thesis, Massachusetts Institute of Technology. (2009).
34. Tutz, G. & Ramzan, S. Improved methods for the imputation of missing data by nearest neighbor methods. *Comput. Stat. Data. Anal.* **90**, 84–99 (2015).
35. Arthur, D. & Vassilvitskii, S. K-means++: The Advantages of Careful Seeding. in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035 (Society for Industrial and Applied Mathematics, Philadelphia, 2007).
36. Jones, R. H. Bayesian information criterion for longitudinal and clustered data. *Stat. Med.* **30**, 3050–3056 (2011).
37. Brown, S. M. et al. Survival after shock requiring high-dose vasopressor therapy. *Chest* **143**, 664–671 (2013).
38. Norris, J. R. Discrete-time Markov chains. in *Markov Chains* (Cambridge University Press, Cambridge, MA, USA, 1997).
39. Jiang, N. & Li, L. *Doubly robust off-policy value evaluation for reinforcement learning*. Preprint at https://arxiv.org/abs/1511.03722 (2015).
40. Thomas, P. S. & Brunskill, E. *Data-efficient off-policy policy evaluation for reinforcement learning*. Preprint at https://arxiv.org/abs/1604.00923 (2016).
41. 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 (Morgan Kaufmann Publishers Inc., Burlington, MA, USA, 2000).
42. Munos, R., Stepleton, T., Harutyunyan, A. & Bellemare, M. G. *Safe and efficient off-policy reinforcement learning*. Preprint at https://arxiv.org/abs/1606.02647 (2016).

nature research

Corresponding author(s): Matthieu Komorowski

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We included all adult patients from two large intensive care databases, MIMIC-III and eICU-RI. We conducted a secondary analysis of patient data initially collected routinely for patient care.

2. Data exclusions

Describe any data exclusions.

Exclusion criteria were pre-established. We excluded patients younger than 18 years old at the time of ICU admission, patients where mortality was not documented and patients with evidence of withdrawal of treatment. In MIMIC-III, we excluded patients where intravenous fluids were not recorded. In eICU-RI, we excluded ICU readmissions and patient admitted in an ICU with insuffcent data collection.

3. Replication

Describe whether the experimental findings were reliably reproduced.

The results were reliably reproduced in a large array of sensitivity analyses, as described in Methods and Extended Data.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not relevant, this was not a randomized controlled study.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not relevant, this was not a randomized controlled study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	□	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	□	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☒	□	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The databases were queried in pgAdmin 4 v 1.2. Computations were implemented in Matlab R2017a (MathWorks Inc., Natick, MA).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used in this study.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used in this study.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used in this study.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used in this study.

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	The study did not involve animals.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The research involved a total of 96,156 unique adult patients admitted to 133 separate intensive care units in the USA. The average age was 65 years old and 53% of the subjects were male. All the patients were diagnosed with sepsis according to the sepsis-3 international definition.