LETTER

https://doi.org/10.1038/s41586-019-1390-1

A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev[1]*, Xavier Glorot[1], Jack W. Rae[1,2], Michal Zielinski[1], Harry Askham[1], Andre Saraiva[1], Anne Mottram[1], Clemens Meyer[1], Suman Ravuri[1], Ivan Protsyuk[1], Alistair Connell[1], Cían O. Hughes[1], Alan Karthikesalingam[1], Julien Cornebise[1,12], Hugh Montgomery[3], Geraint Rees[4], Chris Laing[5], Clifton R. Baker[6], Kelly Peterson[7,8], Ruth Reeves[9], Demis Hassabis[1], Dominic King[1], Mustafa Suleyman[1], Trevor Back[1,13], Christopher Nielson[10,11,13], Joseph R. Ledsam[1,13]* & Shakir Mohamed[1,13]

The early prediction of deterioration could have an important role in supporting healthcare professionals, as an estimated 11% of deaths in hospital follow a failure to promptly recognize and treat deteriorating patients[1]. To achieve this goal requires predictions of patient risk that are continuously updated and accurate, and delivered at an individual level with sufficient context and enough time to act. Here we develop a deep learning approach for the continuous risk prediction of future deterioration in patients, building on recent work that models adverse events from electronic health records[2–17] and using acute kidney injury—a common and potentially life-threatening condition[18]—as an exemplar. Our model was developed on a large, longitudinal dataset of electronic health records that cover diverse clinical environments, comprising 703,782 adult patients across 172 inpatient and 1,062 outpatient sites. Our model predicts 55.8% of all inpatient episodes of acute kidney injury, and 90.2% of all acute kidney injuries that required subsequent administration of dialysis, with a lead time of up to 48 h and a ratio of 2 false alerts for every true alert. In addition to predicting future acute kidney injury, our model provides confidence assessments and a list of the clinical features that are most salient to each prediction, alongside predicted future trajectories for clinically relevant blood tests[9]. Although the recognition and prompt treatment of acute kidney injury is known to be challenging, our approach may offer opportunities for identifying patients at risk within a time window that enables early treatment.

Adverse events and clinical complications are a major cause of mortality and poor outcomes in patients, and substantial effort has been made to improve their recognition[18,19]. Few predictors have found their way into routine clinical practice, because they either lack effective sensitivity and specificity or report damage that already exists[20]. One example relates to acute kidney injury (AKI), a potentially life-threatening condition that affects approximately one in five inpatient admissions in the United States[21]. Although a substantial proportion of cases of AKI are thought to be preventable with early treatment[22], current algorithms for detecting AKI depend on changes in serum creatinine as a marker of acute decline in renal function. Increases in serum creatinine lag behind renal injury by a considerable period, which results in delayed access to treatment. This supports a case for preventative 'screening'-type alerts but there is no evidence that current rule-based alerts improve outcomes[23]. For predictive alerts to be effective, they must empower clinicians to act before a major clinical decline has occurred by: (i) delivering actionable insights on preventable conditions; (ii) being personalized for specific patients; (iii) offering sufficient contextual information to inform clinical decision-making; and (iv) being generally applicable across populations of patients[24].

Promising recent work on modelling adverse events from electronic health records[2–17] suggests that the incorporation of machine learning may enable the early prediction of AKI. Existing examples of sequential AKI risk models have either not demonstrated a clinically applicable level of predictive performance[25] or have focused on predictions across a short time horizon that leaves little time for clinical assessment and intervention[26].

Our proposed system is a recurrent neural network that operates sequentially over individual electronic health records, processing the data one step at a time and building an internal memory that keeps track of relevant information seen up to that point. At each time point, the model outputs a probability of AKI occurring at any stage of severity within the next 48 h (although our approach can be extended to other time windows or severities of AKI; see Extended Data Table 1). When the predicted probability exceeds a specified operating-point threshold, the prediction is considered positive. This model was trained using data that were curated from a multi-site retrospective dataset of 703,782 adult patients from all available sites at the US Department of Veterans Affairs—the largest integrated healthcare system in the United States. The dataset consisted of information that was available from hospital electronic health records in digital format. The total number of independent entries in the dataset was approximately 6 billion, including 620,000 features. Patients were randomized across training (80%), validation (5%), calibration (5%) and test (10%) sets. A ground-truth label for the presence of AKI at any given point in time was added using the internationally accepted 'Kidney Disease: Improving Global Outcomes' (KDIGO) criteria[18]; the incidence of KDIGO AKI was 13.4% of admissions. Detailed descriptions of the model and dataset are provided in the Methods and Extended Data Figs. 1–3.

Figure 1 shows the use of our model. At every point throughout an admission, the model provides updated estimates of future AKI risk along with an associated degree of uncertainty. Providing the uncertainty associated with a prediction may help clinicians to distinguish ambiguous cases from those predictions that are fully supported by the available data. Identifying an increased risk of future AKI sufficiently far in advance is critical, as longer lead times may enable preventative action to be taken. This is possible even when clinicians may not be actively intervening with, or monitoring, a patient. Supplementary Information section A provides more examples of the use of the model.

With our approach, 55.8% of inpatient AKI events of any severity were predicted early, within a window of up to 48 h in advance and with a ratio of 2 false predictions for every true positive. This corresponds to an area under the receiver operating characteristic curve of 92.1%, and an area under the precision–recall curve of 29.7%. When set at this threshold, our predictive model would—if operationalized—trigger a

[1]DeepMind, London, UK. [2]CoMPLEX, Computer Science, University College London, London, UK. [3]Institute for Human Health and Performance, University College London, London, UK. [4]Institute of Cognitive Neuroscience, University College London, London, UK. [5]University College London Hospitals, London, UK. [6]Department of Veterans Affairs, Denver, CO, USA. [7]VA Salt Lake City Healthcare System, Salt Lake City, UT, USA. [8]Division of Epidemiology, University of Utah, Salt Lake City, UT, USA. [9]Department of Veterans Affairs, Nashville, TN, USA. [10]University of Nevada School of Medicine, Reno, NV, USA. [11]Department of Veterans Affairs, Salt Lake City, UT, USA. [12]Present address: University College London, London, UK. [13]These authors contributed equally: Trevor Back, Christopher Nielson, Joseph R. Ledsam, Shakir Mohamed. *e-mail: nenadt@google.com; jledsam@google.com

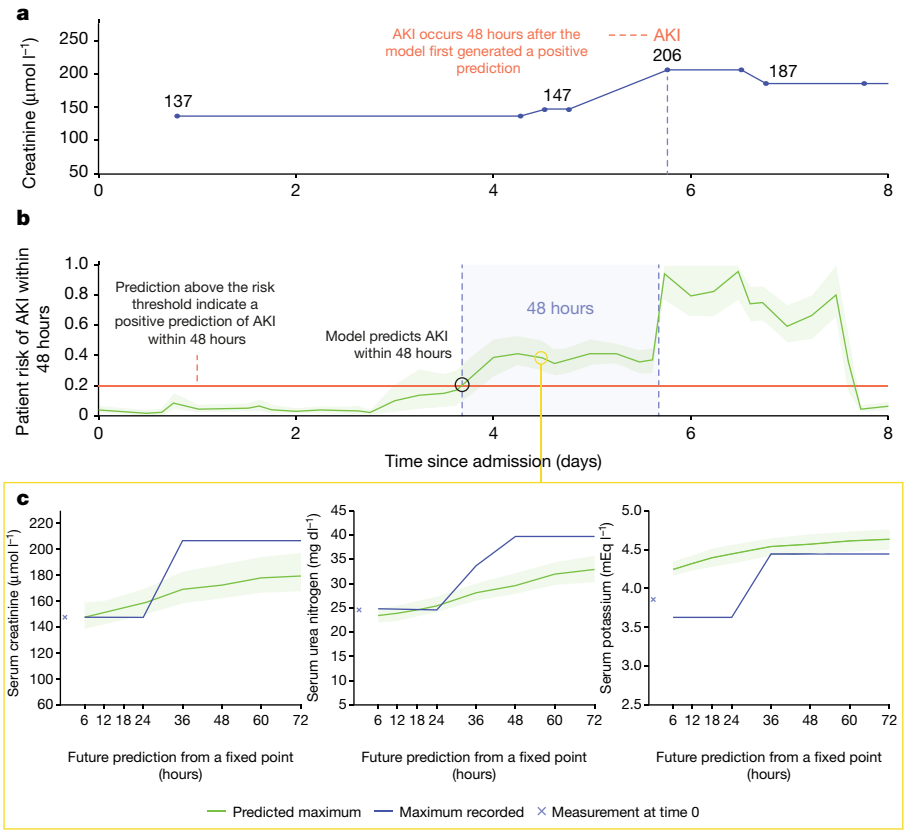


Fig. 1 | Illustrative example of risk prediction, uncertainty and predicted future laboratory values. The first 8 days of admission for a male patient aged 65 with a history of chronic obstructive pulmonary disease. **a**, Patient creatinine measurements during admission. Creatinine measurements, showing AKI occurring on day 5. **b**, Model predictions for any AKI within 48 h. Continuous risk predictions: the model predicted increased AKI risk 48 h before it was observed. A risk above 0.2 (corresponding to 33% precision) was the threshold above which AKI was predicted. Lighter green borders on the risk curve indicate uncertainty, taken as the range of 100 ensemble predictions (after these were trimmed for the highest and lowest 5 values). **c**, Laboratory value predictions 4.5 days into admission. Predictions of the maximum future observed values of creatinine, urea and potassium.

daily clinical assessment in 2.7% of hospitalized patients in this cohort (Extended Data Table 2). Sensitivity was particularly high in patients who went on to develop lasting complications as a result of AKI. The model provided correct early predictions in 84.3% of episodes in which administration of in-hospital or outpatient dialysis was required within 30 days of the onset of AKI of any stage, and in 90.2% of cases in which regular outpatient administration of dialysis was scheduled within 90 days of the onset of AKI (Extended Data Table 3). Figure 2 shows the corresponding receiver operating characteristic and precision–recall curves, as well as a spectrum of operating points of the model. An operating point can be chosen to further increase the proportion of AKI that is predicted early or to reduce the percentage of false predictions at each step, according to clinical priority (Fig. 3). Applied to stage 3 AKI, 84.1% of inpatient events were predicted up to 48 h in advance, with a ratio of 2 false predictions for every true positive (Extended Data Table 4). To respond to these alerts on a daily basis, clinicians would need to attend to approximately 0.8% of in-hospital patients (Extended Data Table 2).

The model correctly identifies substantial future increases in 7 auxiliary biochemical tests in 88.5% of cases (Supplementary Information, section B), and provides information about the factors that are most salient to the computation of each risk prediction. The greatest saliency was identified for laboratory tests that are known to be relevant to renal function (Supplementary Information, section C). The predictive performance of our model was maintained across time and hospital sites, as demonstrated by additional experiments that show generalizability to data acquired at time points after the model was trained (Extended Data Table 5).

Our approach significantly outperformed ($P < 0.001$) established state-of-the-art baseline models (Supplementary Information, section D). For example, we implemented a baseline model with gradient-boosted trees using manually curated features that are known to be relevant for modelling kidney function and in the delivery of routine care (Supplementary Information, sections E and F), combined with aggregate statistical information on trends observed in the recent history of the patient. This yielded 3,599 clinically relevant features that were provided to the baseline at each step (Methods). For the same level of precision, this baseline model was able to detect 36.0% of all episodes of AKI in inpatients up to 48 h in advance, compared to 55.8% for our model.

Of the false-positive alerts made by our model, 24.9% were positive predictions that were made even earlier than the 48-h window in patients who subsequently developed AKI (Extended Data Fig. 4). Of these, 57.1% occurred in patients with pre-existing chronic kidney disease, who are at a higher risk of developing AKI. Of the remaining false-positive alerts, 24.1% were trailing predictions that occurred after an AKI episode appeared to have resolved; alerts such as these can be filtered out in clinical practice. For positive risk predictions in which no AKI was subsequently observed (in this retrospective dataset), it is probable that many occurred in patients at risk of AKI to whom appropriate preventative treatment was administered—which would have averted subsequent AKI. In addition to these early and trailing predictions, 88% of the remaining false-positive alerts occurred in patients with severe renal impairment, known renal pathology or evidence in the electronic health record that the patient required clinical review (Extended Data Fig. 4).

Our aim is to provide risk predictions that enable personalized preventative action to be delivered at a large scale. The way these predictions are used may vary by clinical setting: a trainee doctor could be

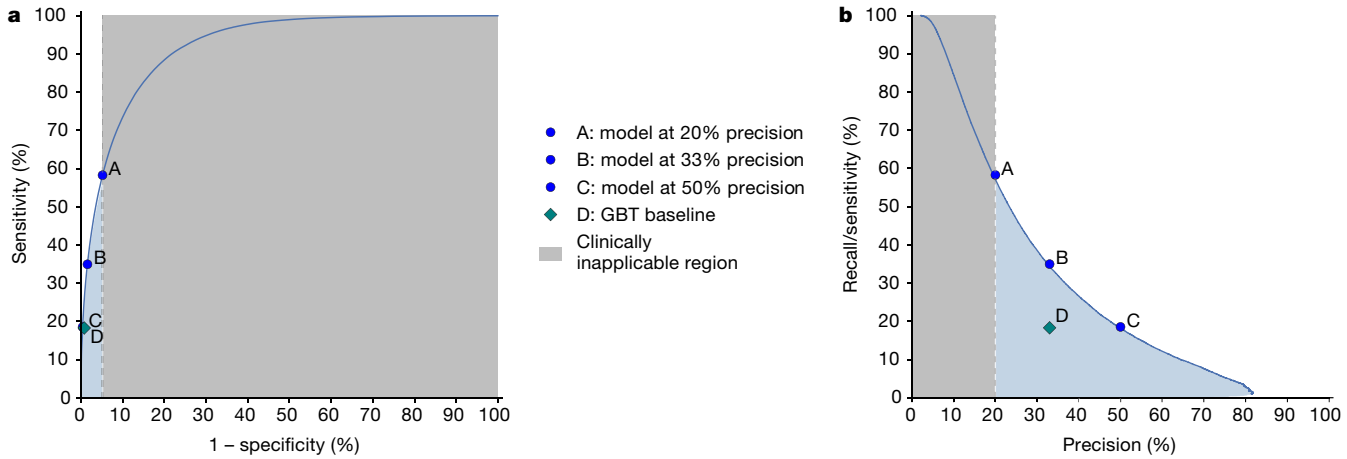


Fig. 2 | Model performance illustrated by receiver operating characteristic and precision–recall curves. a, **b**, Receiver operating characteristic (**a**) and precision–recall (**b**) curves for the risk that AKI of any severity will occur within 48 h. Blue dots represent different model operating points (see Extended Data Table 4). Grey shaded area corresponds to operating points with more than four false positives for each true positive. Blue shaded area represents performance in the part of the operating space that is more clinically applicable. The model significantly ($P < 1 \times 10^{-6}$) outperformed the gradient-boosted tree baseline, shown in **b** for operating-point C using two-sided Mann–Whitney *U*-test on 200 samples per model (see Methods).

alerted in real time to each patient under their care, and specialist nephrologists or rapid-response teams[27] could identify high-risk patients to prioritize their response. This is possible because performance was consistent across multiple clinically important groups—notably, those at an increased risk of AKI (Supplementary Information, section G). Our model is designed to complement existing routine care, as it is trained specifically to predict episodes of AKI that happened in this retrospective dataset despite existing best practices.

Although we demonstrate a model that is trained and evaluated on a clinically representative set of patients from the entire US Department of Veterans Affairs healthcare system, this demographic is not representative of the global population. Female patients comprised 6.38% of patients in the dataset, and model performance was lower for this demographic (Extended Data Table 6). Validating the predictive performance of the proposed system on a general population would require training and evaluating the model on additional representative datasets. Future work will need to address the under-representation of sub-populations in the training data[28] and overcome the effect of potential confounding factors that relate to hospital processes[29]. KDIGO is an indicator of AKI that has a long lag time after the initial renal impairment, and model performance could be enhanced by improvements in the ground-truth definition of AKI and in data quality[30].

Despite the state-of-the-art retrospective performance of our model compared to existing literature, future work should now prospectively evaluate and independently validate the proposed model to establish its clinical utility and effect on patient outcomes, as well as explore the role of the model in researching strategies for delivering preventative care for AKI.

In summary, we demonstrate a deep learning approach for the continuous prediction of AKI within a clinically actionable window of up to 48 h in advance. We report performance on a clinically diverse population and across a large number of sites to show that our approach may allow for the delivery of potentially preventative treatment—before the physiological insult itself, in a large number of the cases. Our results open up the possibility for deep learning to guide the prevention of clinically important adverse events. With the possibility of risk predictions delivered in clinically actionable windows, alongside the increasing size and scope of electronic health record datasets, we now shift to a regime in which the role of machine learning in clinical care can grow rapidly, supplying tools for enhancing patient and clinician experiences and

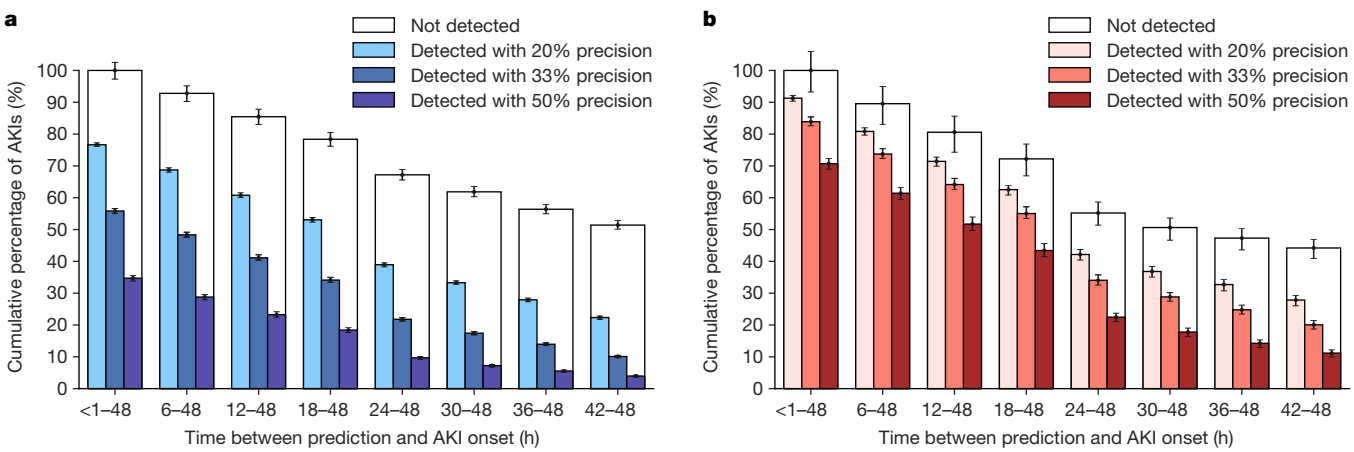


Fig. 3 | The time between model prediction and the actual AKI event. The models predict AKI risk within a particular time window. Within this, the time in hours between prediction and AKI can vary (error bars, bootstrap pivotal 95% confidence intervals; $n = 200$). **a**, **b**, Prediction performance for any AKI (**a**) and AKI stage 3 (**b**) 48 h ahead of time, shown for different precisions. A greater proportion of AKI events were correctly predicted closer to the time step immediately before the AKI. The available time window for prediction is shortened in AKI events which occur less than 48 h after admission. For each column, the boxed area shows the upper limit on possible predictions.

potentially becoming a ubiquitous and integral part of routine clinical pathways.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1390-1.

Received: 13 November 2018; Accepted: 18 June 2019;
Published online 31 July 2019.

1. Thomson, R., Luettel, D., Healey, F. & Scobie, S. *Safer Care for the Acutely Ill Patient: Learning from Serious Incidents* (National Patient Safety Agency, 2007).
2. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWscore) for septic shock. *Sci. Transl. Med.* **7**, 299ra122 (2015).
3. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
4. Koyner, J. L., Adhikari, R., Edelson, D. P. & Churpek, M. M. Development of a multicenter ward-based AKI prediction model. *Clin. J. Am. Soc. Nephrol.* **11**, 1935–1943 (2016).
5. Cheng, P., Waitman, L. R., Hu, Y. & Liu, M. Predicting inpatient acute kidney injury over different time horizons: how early and accurate? In *AMIA Annual Symposium Proceedings* 565 (American Medical Informatics Association, 2017).
6. Koyner, J. L., Carey, K. A., Edelson, D. P. & Churpek, M. M. The development of a machine learning inpatient acute kidney injury prediction model. *Crit. Care Med.* **46**, 1070–1077 (2018).
7. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
8. Avati, A. et al. Improving palliative care with deep learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 311–316 (2017).
9. Lim, B. & van der Schaar, M. Disease-Atlas: navigating disease trajectories with deep learning. *Proc. Mach. Learn. Res.* **85**, 137–160 (2018).
10. Futoma, J., Hariharan, S. & Heller, K. A. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *Proc. International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 1174–1182 (2017).
11. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).
12. Lipton, Z. C., Kale, D. C., Elkan, C. & Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. Preprint at https://arxiv.org/abs/1511.03677 (2016).
13. Cheng, Y. P. Z. J. H. & Wang, F. Risk prediction with electronic health records: a deep learning approach. In *Proc. SIAM International Conference on Data Mining* (eds Venkatasubramanian, S. C. & Meria, W.) 432–440 (2016).
14. Soleimani, H., Subbaswamy, A. & Saria, S. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In *Proc. 33rd Conference on Uncertainty in Artificial Intelligence* (AUAI Press Corvallis, 2017).
15. Alaa, A. M., Yoon, J., Hu, S. & van der Schaar, M. Personalized risk scoring for critical care prognosis using mixtures of Gaussian process experts. *IEEE Trans. Biomed. Eng.* **65**, 207–218 (2018).
16. Perotte, A., Ranganath, R., Hirsch, J. S., Blei, D. & Elhadad, N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *J. Am. Med. Inform. Assoc.* **22**, 872–880 (2015).
17. Bihorac, A. et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann. Surg.* **269**, 652–662 (2019).
18. Khwaja, A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin. Pract.* **120**, c179–c184 (2012).
19. Stenhouse, C., Coates, S., Tivey, M., Allsop, P. & Parker, T. Prospective evaluation of a modified early warning score to aid earlier detection of patients developing critical illness on a general surgical ward. *Br. J. Anaesth.* **84**, 663P (2000).
20. Alge, J. L. & Arthur, J. M. Biomarkers of AKI: a review of mechanistic relevance and potential therapeutic implications. *Clin. J. Am. Soc. Nephrol.* **10**, 147–155 (2015).
21. Wang, H. E., Muntner, P., Chertow, G. M. & Warnock, D. G. Acute kidney injury and mortality in hospitalized patients. *Am. J. Nephrol.* **35**, 349–355 (2012).
22. MacLeod, A. NCEPOD report on acute kidney injury—must do better. *Lancet* **374**, 1405–1406 (2009).
23. Lachance, P. et al. Association between e-alert implementation for detection of acute kidney injury and outcomes: a systematic review. *Nephrol. Dial. Transplant.* **32**, 265–272 (2017).
24. Johnson, A. E. W. et al. Machine learning and decision support in critical care. *Proc. IEEE Inst. Electr. Electron Eng.* **104**, 444–466 (2016).
25. Mohamadlou, H. et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can. J. Kidney Health Dis.* **5**, 1–9 (2018).
26. Pan, Z. et al. A self-correcting deep learning approach to predict acute conditions in critical care. Preprint at https://arxiv.org/abs/1901.04364 (2019).
27. Park, S. et al. Impact of electronic acute kidney injury (AKI) alerts with automated nephrologist consultation on detection and severity of AKI: a quality improvement study. *Am. J. Kidney Dis.* **71**, 9–19 (2018).
28. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? Preprint at https://arxiv.org/abs/1805.12002 (2018).
29. Schulam, P. & Saria, S. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 1697–1708 (2017).
30. Telenti, A., Steinhubl, S. R. & Topol, E. J. Rethinking the medical record. *Lancet* **391**, 1013 (2018).

METHODS

Data description. The clinical data used in this study were collected by the US Department of Veterans Affairs and transferred to DeepMind in a de-identified format. No personal information was included in the dataset, which met HIPAA 'Safe Harbor' criteria for de-identification.

The US Department of Veterans Affairs serves a population of over nine million veterans and their families across the entire United States of America. The US Department of Veterans Affairs is composed of 1,243 healthcare facilities (sites), including 172 Veterans Affairs Medical Centers and 1,062 outpatient facilities[31]. Data from these sites are aggregated into 130 data centres, of which 114 had data from inpatient admissions that we used in this study. Four sites were excluded because they had fewer than 250 admissions during the 5-year time period. No other patients were excluded based on location.

The data comprised all patients aged between 18 and 90 who were admitted for secondary care to medical or surgical services from the beginning of October 2011 to the end of September 2015 (including laboratory data) and for whom there was at least one year of electronic health record data before admission. The data included medical records with entries up to ten years before each admission date and up to two years afterwards, where available. Where available in the US Department of Veterans Affairs database, data included outpatient visits, admissions, diagnoses as International Statistical Classification of Diseases and Related Health Problems (ICD9) codes, procedures as Current Procedural Terminology (CPT) codes, laboratory results (including—but not limited to—biochemistry, haematology, cytology, toxicology, microbiology and histopathology), medications and prescriptions, orders, vital signs, health factors and note titles. Free text and diagnoses that were rare (fewer than 12 distinct patients with at least 1 occurrence in the US Department of Veterans Affairs database) were excluded to ensure all potential privacy concerns were addressed. In addition, conditions that were considered sensitive were excluded before transfer, such as patients with HIV/AIDS, sexually transmitted diseases, substance abuse and those admitted to mental health services.

The final dataset consisted of all eligible patients following this set of inclusion criteria, and comprised 703,782 patients, which provided 6,352,945,637 clinical-event entries. Each clinical entry denoted a single procedure, laboratory test result, prescription, diagnosis and so on: 3,958,637,494 entries came from outpatient events and the remaining 2,394,308,143 events came from admissions. Extended Data Table 6 contains an overview of patient demographics in the data as well as the prevalence of conditions that are associated with AKI across the data splits. The final dataset was randomly divided into training (80% of observations), validation (5%), calibration (5%) and testing (10%) sets. All data for a single patient were assigned to exactly one of these splits. The test population consisted of 70,681 individual patients and 252,492 unique admissions. A sample size of 179 patients would be required to detect sensitivity and specificity at 0.05 marginal error and 95% confidence. When assigning patients randomly to test, calibration, validation and training groups investigators were blinded to patient covariates and all features in the electronic health record that were not required to perform the research (for example, creatinine was required to label AKI as a ground truth). Patient recruitment was conducted by independent members of the Department of Veterans Affairs National Data Center; research team members were blinded to this recruitment.

Data preprocessing. *Feature representation.* Every patient in the dataset was represented by a sequence of events, with each event providing the patient information that was recorded within a six-hour period; that is, each day was broken into four six-hour periods and all records that occurred within the same six-hour period were grouped together. The available data within these six-hour windows, along with additional summary statistics and augmentations, formed a feature set that was used as input to our predictive models. Extended Data Figure 1 provides a diagrammatic view of a patient sequence and its temporal structure.

We did not perform any imputation of missing numerical values, because explicit imputation of missing values does not always provide consistent improvements to predictive models based on electronic health records[32]. Instead, we associated each numerical feature with one or more discrete 'presence' features to enable our models to distinguish between the absence of a numerical value and an actual value of zero. Additionally, these presence features encoded whether a particular numerical value is considered to be normal, low, high, very low or very high. For some data points, the explicit numerical values were not recorded (usually when the values were considered normal), and the provision of this encoding of the numerical data allowed our models to process these measurements even in their absence. Discrete features, such as diagnostics or procedural codes, were also encoded as binary presence features.

All numerical features were normalized to the [0, 1] range after capping the extreme values at the 1st and 99th percentile. This prevents the normalization from being dominated by potentially large data entry errors, while preserving most of the signal.

Each clinical feature was mapped onto a corresponding high-level concept, such as procedure, diagnosis, prescription, laboratory test, vital sign, admission, transfer and so on. In total, 29 such high-level concepts were present in the data. At each step, a histogram of the frequencies of these concepts among the clinical entries that took place at that step was provided to the models, along with the numerical and binary presence features.

The approximate age of each patient in days (as well as the six-hour period in the day to which the data were associated) were provided as explicit features to the models. In addition, we provided some simple features that made it easier for the models to predict the risk of developing AKI. In particular, we provided the median yearly creatinine baseline and the minimum 48-h creatinine baseline as additional numerical features. These are the baseline values that are used in the KDIGO criteria and help to give important context to the models regarding how to interpret new serum creatinine measurements as they become available.

We additionally computed 3 historical aggregate feature representations at each step: one for the past 48 h, one for the past 6 months and one for the past 5 years. All histories were optionally provided to the models and the decision on which combination of historical data to include was based on the model performance on the validation set. We did this historical aggregation for discrete features by including whether these features were observed in the historical interval or not. For numerical features, we included the count, mean, median, standard deviation, minimum and maximum value observed in the interval, as well as simple trend features such as the difference between the last observed value and the minimum or maximum and the average difference between subsequent steps (which measures the temporal short-term variability of the measurement). Supplementary Information section H provides the effect of volume and recency of available data on model performance.

Because patient measurements are made irregularly, not all six-hour time periods in a day will have new data associated with them. Our models operate at regular time intervals regardless, and all time periods without new measurements include only the available metadata and (optionally) the historical aggregate features. This approach makes continuous risk predictions possible, and allows our models to use patterns of missingness in the data during the training process.

For about 35% of all entries, the day—but not the specific time during the day—on which they occurred was known. For each day in the sequence of events, we aggregated these unknown-time entries into a specific bucket that was appended to the end of the day. This ensured that our models could iterate over this information without potentially leaking information from the future. Our models were not allowed to make predictions from these surrogate points and they were not factored into the evaluation. The models can use the information contained within the surrogate points on the next time step, corresponding to the first interval of the following day.

Diagnoses in the data are sometimes known to be recorded in the electronic health record before the time when an actual diagnosis was made clinically. To avoid leaking future information to the models, we shifted all of the diagnoses within each admission to the very end of that admission and only provided them to the models at that point, at which time they can be factored in for future admissions. This discards potentially useful information: the performance obtained in this way is conservative by design and it is possible that, in reality, the models would be able to perform better with this information provided in a consistent way.

Ground-truth labels using KDIGO. The patient AKI states were computed at each time step on the basis of the KDIGO[18] criteria, the recommendations of which are based on systematic reviews of relevant trials. KDIGO accepts three definitions of AKI[18]: an increase in serum creatinine of 0.3 mg/dl (26.5 μmol/l) within 48 h; an increase in serum creatinine of 1.5× the baseline creatinine level of a patient, known or presumed to have occurred within the previous 7 days; or a urine output of <0.5 ml/kg/h over 6 h. The first two definitions were used to provide ground-truth labels for the onset of AKI; the third definition could not be used, as urine output was not recorded digitally in the majority of sites that formed part of this work. A baseline of median annualized creatinine was used when previous measurements were available; when these measurements were not present the 'modification of diet in renal disease' formula was applied to estimate baseline creatinine. Using the KDIGO criteria based on serum creatinine and its corresponding definitions for AKI severity, three AKI categories were obtained: 'all AKI' (KDIGO stages 1, 2 and 3), 'moderate and severe AKI' (KDIGO stages 2 and 3), and 'severe AKI' (KDIGO stage 3).

The AKI stages were computed at times at which there was a serum creatinine measurement present in the sequence, and then copied forward in time until the next creatinine measurement, at which time the ground-truth AKI state was updated accordingly. To avoid basing the current estimate of the KDIGO AKI stage on a previous measurement that may no longer be reliable, the AKI states were propagated for (at most) four days forward in case no new creatinine measurements were observed. From that point onwards, AKI states were marked as unknown. Patients who experience AKI tend to be closely monitored and their levels of serum creatinine are measured regularly, so an absence of a measurement

for multiple days in such cases is uncommon. A gap of 4 days between subsequent creatinine measurements represents the 95th percentile in the distribution of time between 2 consecutive creatinine measurements.

The prediction target at each point in time is a binary variable that is positive if the AKI category of interest (for example, all AKI) occurs within a chosen future time horizon. If no AKI state was recorded within the chosen horizon, this was interpreted as a negative. We used 8 future time horizons (6-h, 12-h, 18-h, 24-h, 36-h, 48-h, 60-h and 72-h ahead) that were all available at each time point.

Event sequences of patients who were undergoing renal replacement therapy were excluded from the target labels heuristically (on the basis of data entries for renal replacement therapy procedures being performed being present in the electronic health record), for the duration of dialysis administration. We have excluded entire sub-sequences of events between entries for renal replacement therapy procedures that occur within a week of each other. The edges of the sub-sequence were also appropriately excluded from label computations.

Models for predicting AKI. Our predictive system operates sequentially over the electronic health record. At each time point, input features (as described in 'Feature representation') were provided to a statistical model, the output of which is a probability of any-severity stage of AKI occurring in the next 48 h. If this probability exceeds a chosen operating threshold, we make a positive prediction that can then trigger an alert. This is a general framework within which existing approaches also fit, and we describe the baseline methods in 'Competitive baseline methods' below. The contribution of this work is in the design of the particular model that is used and its training procedure, and the demonstration of its effectiveness—on a large-scale electronic health record dataset and across many different regimes— in making useful predictions of future AKI.

Extended Data Figure 2 gives a schematic view of our model, which makes predictions by first transforming the input features using an embedding module. This embedding is fed into a multi-layer recurrent neural network, the output of which at every time point is fed into a prediction module that provides the probability of future AKI at the time horizon for which the model will be trained. The entire model can be trained end-to-end; that is, the parameters can be learned jointly without pre-training any parts of the model. To provide useful predictions, we train an ensemble of predictors to estimate the confidence of the model, and the resulting ensemble predictions are then calibrated using isotonic regression to reflect the frequency of observed outcomes[33].

Embedding modules. The embedding layers transform the high-dimensional and sparse input features into a lower-dimensional continuous representation that makes subsequent prediction easier. We use a deep multilayer perceptron with residual connections and rectified-linear activations. We use L_1 regularization on the embedding parameters to prevent overfitting and to ensure that our model focuses on the most-salient features. We compared simpler linear transformations, which did not perform as well as the multi-layer version we used. We also compared unsupervised approaches such as factor analysis, standard auto-encoders and variational auto-encoders, but did not find any substantial advantages in using these methods.

Recurrent neural network core. Recurrent neural networks (RNNs) run sequentially over the electronic health record entries and are able to implicitly model the historical context of a patient by modifying an internal representation (or state) through time. We use a stacked multiple-layer recurrent network with highway connections between each layer[34], which at each time step takes the embedding vector as an input. We use the simple recurrent unit network as the RNN architecture, with tanh activations. We chose this from a broad range of alternative RNN architectures: specifically, the long short-term memory[35], update gate RNN and intersection RNN[36], simple recurrent units[37,38], gated recurrent units[39], the neural Turing machine[40], memory-augmented neural network[41], the Differentiable Neural Computer[42] and the relational memory core[43]. These alternatives did not provide significant performance improvements over the simple recurrent unit architecture (Supplementary Information section D).

Prediction targets and training objectives. The output of the RNN is fed to a final linear prediction layer that makes predictions over all eight future prediction windows (6-h windows from 6 h ahead to 72 h ahead). We use a cumulative distribution function layer across time windows to encourage monotonicity, because the presence of AKI within a shorter time window implies a presence of AKI within a longer time window. Each of the resulting eight outputs provides a binary prediction for AKI severity at a specific time window and is compared to the ground-truth label using the cross-entropy loss function (Bernoulli log-likelihood).

We also make a set of auxiliary numerical predictions; at each step, we also predict the maximum future observed value of a set of laboratory tests over the same set of time intervals as we use to make the future AKI predictions. The laboratory tests predicted are ones that are known to be relevant to kidney function: specifically, creatinine, urea nitrogen, sodium, potassium, chloride, calcium and phosphate. This multi-task approach results in better generalization and more-robust representations, especially under class imbalance[44–46]. The overall improvement that we observed from including the auxiliary task was around 3% area under the precision–recall curve in most cases (see Supplementary Information section A for more details).

Our overall loss function is the weighted sum of the cross-entropy loss from the AKI predictions and the squared loss for each of the seven laboratory-test predictions. We investigated the use of oversampling and overweighting of the positive labels to account for class imbalance. For oversampling, each mini-batch contains a larger percentage of positive samples than average in the entire dataset. For overweighting, the prediction for positive labels contributes proportionally more to the total loss.

Training and hyperparameters. We selected our proposed model architecture among several alternatives on the basis of the validation set performance (Supplementary Information section D), and subsequently performed an ablation analysis of the design choices (Supplementary Information section I). All variables are initialized via normalized (Xavier) initialization[47] and trained using the Adam optimization scheme[48]. We use exponential learning-rate decay during training. The best validation results were achieved using an initial learning rate of 0.001 decayed every 12,000 training steps by a factor of 0.85, with a batch size of 128 and a back-propagation through time window of 128. The embedding layer is of size 400 for each of the numerical and presence input features (800 in total when concatenated) and uses 2 layers. The best-performing RNN architecture used a cell size of 200 units per layer and 3 layers. A detailed overview of different hyperparameter combinations evaluated in the experiments is available in Supplementary Information section J. We conducted extensive hyperparameter explorations of dropout rates for different kinds of dropout to determine the best model regularization. We have considered input dropout, output dropout, embedding dropout, cell-state dropout and variational dropout. None of these led to improvements, so dropout is not included in our model.

Competitive baseline methods. Established models for future AKI prediction make use of L_1-regularized logistic regression or gradient-boosted trees, trained on a clinically relevant set of features that are known to be important either for routine clinical practice or the modelling of kidney function. A curated set of clinically relevant features was chosen using existing AKI literature (Supplementary Information section F) and the consensus opinion of six clinicians: three senior attending physicians with over twenty years expertise, one nephrologist and two intensive care specialists; and three clinical residents with expertise in nephrology, internal medicine and surgery. This set was further extended to include 36 of the most-salient features discovered by our deep learning model that were not in the original list, to give further predictive signal to the baseline. The final curated dataset contained 315 base features of demographics, admission information, vital sign measurements, select laboratory tests and medications, and diagnoses of chronic conditions that are directly associated with an increased risk of AKI. The full feature set is listed in Supplementary Information section E. We additionally computed a set of manually engineered features (yearly and 48-hourly baseline creatinine levels (consistent with KDIGO guidelines), the ratio of blood urea nitrogen to serum creatinine, grouped severely reduced glomerular filtration rate (corresponding to stages 3a to 5), and flagging patients with diabetes by combining ICD9 codes and values of measured haemoglobin A1c) and a representation of the short-term and long-term history of a patient (see 'Feature representation'). These features were provided explicitly, because the interaction terms and historical trends might not have been recovered by simpler models. This resulted in a total of 3,599 possible features for the baseline model. We provide a table with a full set of baseline comparisons in Supplementary Information section D.

Evaluation. The data were split into training, validation, calibration and test sets in such a way that information from a given patient was present only in one split. The training split was used to train the proposed models. The validation set was used to iteratively improve the models by selecting the best model architectures and hyperparameters.

The models selected on the validation set were recalibrated on the calibration set in order to further improve the quality of the risk predictions. Deep learning models with softmax or sigmoid output trained with cross-entropy loss are prone to miscalibration, and recalibration ensures that consistent probabilistic interpretations of the model predictions can be made[49]. For calibration, we considered Platt scaling[50] and isotonic regression[33]. To compare uncalibrated predictions to recalibrated ones, we used the Brier score[51] and reliability plots[52]. The best models were evaluated on the independent test set that was retained during model development.

The main metrics used in model selection and the final report are: the AKI episode sensitivity, the area under the precision–recall curve, the area under the receiver operating curve, and the per-step precision, per-step sensitivity and per-step specificity. The AKI episode sensitivity corresponds to the percentage of all AKI episodes that were correctly predicted ahead of time within the corresponding time windows of up to 48 h. By contrast, the precision is computed per step because the predictions are made at each step, to account for the rate of false alerts over time.

Owing to the sequential nature of making predictions, the total number of positive steps does not directly correspond to the total number of distinct AKI episodes. Multiple positive alerting opportunities may be associated with a single AKI episode and AKI episodes may offer a number of such early alerting steps, depending on how late they occur within the admission. AKI episodes that occur later during in-hospital stay can be predicted earlier than an AKI episode that occurs immediately upon admission. To better assess the clinical applicability of the proposed model, we explicitly compute the AKI episode sensitivity for different levels of stepwise precision.

Given that the models were designed for continuous monitoring and risk prediction, they were evaluated at each 6-h time step within all of the admissions for each patient except for the steps within AKI episodes (which were ignored). The models were not evaluated on outpatient events. All steps for which there was no record of AKI occurring in the relevant future time window were considered to be negative examples.

Approximately 2% of individual time steps presented to the models sequentially were associated with a positive AKI label, so the AKI prediction task is class-imbalanced. For per-step performance metrics, we report both the area under the receiver operating characteristic curve as well as the area under the precision–recall curve. Area under the precision–recall curve is known to be more informative for class-imbalanced predictive tasks[53], as it is more sensitive to changes in the number of false-positive predictions.

To gauge uncertainty on the performance of a trained model, we calculated 95% confidence intervals with the pivot bootstrap estimator[54]. This was done by sampling the entire validation and test dataset with replacement 200 times. Because bootstrapping assumes the resampling of independent events, we resample entire patients instead of resampling individual admissions or time steps. Where appropriate, we also compute a two-sided Mann–Whitney U-test[55] on the samples for the respective models.

To quantify the uncertainty on model predictions (versus overall performance), we trained an ensemble of 100 models with a fixed set of hyperparameters but different initial seeds. This follows similar uncertainty approaches in supervised learning[56] and medical imaging predictions[57]. The prediction confidence was assessed by inspecting the variance over the 100 model predictions from the ensemble. This confidence reflected the accuracy of a prediction: the mean standard deviation of false-positive predictions was higher than the mean standard deviation of true-positive predictions and similarly for false-negative versus true-negative predictions ($P < 0.01$) (Supplementary Information section K).

Ethics and information governance. This work, and the collection of data on implied consent, received Tennessee Valley Healthcare System Institutional Review Board (IRB) approval from the US Department of Veterans Affairs. De-identification was performed in line with the Health Insurance Portability and Accountability Act (HIPAA), and validated by the US Department of Veterans Affairs Central Database and Information Governance departments. Only de-identified retrospective data were used for research, without the active involvement of patients.

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

Data availability

The clinical data used for the training, validation and test sets were collected at the US Department of Veterans Affairs and transferred to a secure data centre with strict access controls in de-identified format. Data were used with both local and national permissions. It is not publicly available and restrictions apply to its use. The de-identified dataset (or a test subset) may be available from the US Department of Veterans Affairs, subject to local and national ethical approvals.

Code availability

We make use of several open-source libraries to conduct our experiments: the machine learning framework TensorFlow (https://github.com/tensorflow/tensorflow) along with the TensorFlow library Sonnet (https://github.com/deepmind/sonnet), which provides implementations of individual model components[58]. Our experimental framework makes use of proprietary libraries and we are unable to publicly release this code. We detail the experiments and implementation details in the Methods and Supplementary Information to allow for independent replication.

31. Department of Veterans Affairs. *Veterans Health Administration: Providing Health Care for Veterans*. https://www.va.gov/health/ (accessed 9 November 2018).
32. Razavian, N. & Sontag, D. Temporal convolutional neural networks for diagnosis from lab tests. In *4th Int. Conf. Learn. Representations* (2016).
33. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds, Zaïane, O. R. et al.) 694–699 (ACM, 2002).
34. Zilly, J. G., Srivastava, R. K., Koutník, J. & Schmidhuber, J. Recurrent highway networks. In *Proc. International Conference on Machine Learning (vol. 70)* (eds Precup, D. & Teh, Y. W.) 4189–4198 (2017).
35. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
36. Collins, J., Sohl-Dickstein, J. & Sussillo, D. Capacity and trainability in recurrent neural networks. In *International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) https://openreview.net/forum?id=BydARw9ex (2017).
37. Bradbury, J., Merity, S., Xiong, C. & Socher, R. Quasi-recurrent neural networks. In *International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) https://openreview.net/forum?id=H1zJ-v5xl (2017).
38. Lei, T. & Zhang, Y. Training RNNs as fast as CNNs. Preprint at https://arxiv.org/abs/1709.02755v1 (2017).
39. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modelling. Preprint at https://arxiv.org/abs/1412.3555 (2014).
40. Graves, A., Wayne, G. & Danihelka, I. Neural Turing machines. Preprint at https://arxiv.org/abs/1410.5401 (2014).
41. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. Meta-learning with memory-augmented neural networks. In *Proc. International Conference on Machine Learning* (eds Balcan, M. F. & Weinberger, K. Q.) 1842–1850 (2016).
42. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
43. Santoro, A. et al. Relational recurrent neural networks. In *Advances in Neural Information Processing Systems 31* (eds Bengio, S. et al.) 7310–7321 (2018).
44. Caruana, R., Baluja, S. & Mitchell, T. in *Advances in Neural Information Processing Systems* (eds Mozer, M. et al.) 959–965 (1996).
45. Wiens, J., Guttag, J. & Horvitz, E. Patient risk stratification with time-varying parameters: a multitask learning approach. *J. Mach. Learn. Res.* **17**, 1–23 (2016).
46. Ding, D. Y. et al. The effectiveness of multitask learning for phenotyping with electronic health records data. Preprint at https://arxiv.org/abs/1808.03331v1 (2018).
47. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (vol. 9)* (eds Tehand, Y. W. & Titterington, M.) 249–256 (2010).
48. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) https://dblp.org/rec/bib/journals/corr/KingmaB14 (2015).
49. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proc. International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 1321–1330 (2017).
50. Platt, J. C. in *Advances in Large-Margin Classifiers* (eds Smola, A. et al.) 61–74 (MIT Press, 1999).
51. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weath. Rev.* **78**, 1–3 (1950).
52. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proc. International Conference on Machine Learning* (eds Raedt, L. D. & Wrobel, S.) 625–632 (ACM, 2005).
53. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
54. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC, 1994).
55. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
56. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) 6402–6413 (2017).
57. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
58. Abadi, M. et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467 (2015).

Acknowledgements We thank the veterans and their families under the care of the US Department of Veterans Affairs. We thank A. Graves, O. Vinyals, K. Kavukcuoglu, S. Chiappa, T. Lillicrap, R. Raine, P. Keane, M. Seneviratne, A. Schlosberg, O. Ronneberger, J. De Fauw, K. Ruark, M. Jones, J. Quinn, D. Chou, C. Meaden, G. Screen, W. West, R. West, P. Sundberg and the Google AI team, J. Besley, M. Bawn, K. Ayoub and R. Ahmed. Finally, we thank the many physicians, administrators and researchers of the US Department of Veterans Affairs who worked on the data collection, and the rest of the DeepMind team for their support, ideas and encouragement. G.R. and H.M. were supported by University College London and the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre. The views expressed are those of these author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Author contributions M.S., T.B., J.C., J.R.L., N.T., C.N., D.H. and R.R. initiated the project. N.T., X.G., H.A., A.S., J.R.L., C.N., C.R.B. and K.P. created the dataset. N.T., X.G., A.S., H.A., J.W.R., M.Z., A.M., I.P. and S.M. contributed to software engineering. N.T., X.G., A.M., J.W.R., M.Z., A.S., C.B., S.M., J.R.L. and C.N. analysed the results. N.T., X.G., A.M., J.W.R., M.Z., A.S., H.A., J.C., C.O.H., C.R.B., T.B., C.N., S.M. and J.R.L. contributed to the overall experimental design. N.T., X.G., A.M.,

J.W.R., M.Z., S.R. and S.M. designed the model architectures. J.R.L., G.R., H.M., C.L., A.C., A.K., C.O.H., D.K. and C.N. contributed clinical expertise. A.M., N.T., M.Z. and J.W.R. contributed to experiments into model confidence. M.Z., N.T., A.S., A.M. and J.W.R. contributed to model calibration. N.T., M.Z., A.M., A.S., X.G. and J.R.L. contributed to false-positive analysis. N.T., X.G., A.M., J.W.R., M.Z., A.S., S.R. and S.M. contributed to comparison of different architectures. N.T., A.M., X.G., A.S., M.Z., J.R.L. and S.M. contributed to experiments on auxiliary prediction targets. A.M., N.T., X.G., M.Z., A.S., J.R.L. and S.M. contributed to experiments into model generalizability. M.Z., A.M., N.T., T.B. and J.R.L. contributed to subgroup analyses. J.W.R., N.T., A.S., M.Z. and S.M. contributed to ablation experiments. N.T., A.S. and J.R.L. contributed to experiments into how to handle renal replacement therapy in the data. J.W.R., X.G., N.T., A.M., A.C., C.N., K.P., C.R.B., M.Z., A.S. and J.R.L. contributed to analysing salient clinical features. A.M., M.Z. and N.T. contributed to experiments into the influence of data recency on model performance. C.M., S.M., H.A., C.N., J.R.L. and T.B. managed the project. N.T., J.R.L., J.W.R., M.Z., A.M., H.M., C.R.B., S.M. and G.R. wrote the paper.

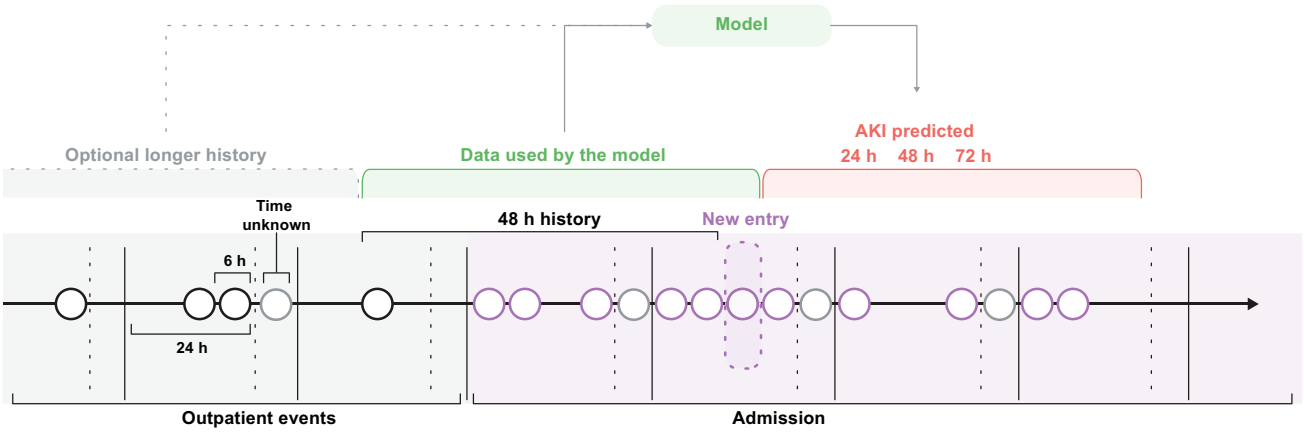
Competing interests G.R., H.M. and C.L. are paid contractors of DeepMind. The authors have no other competing interests to disclose.

Additional information

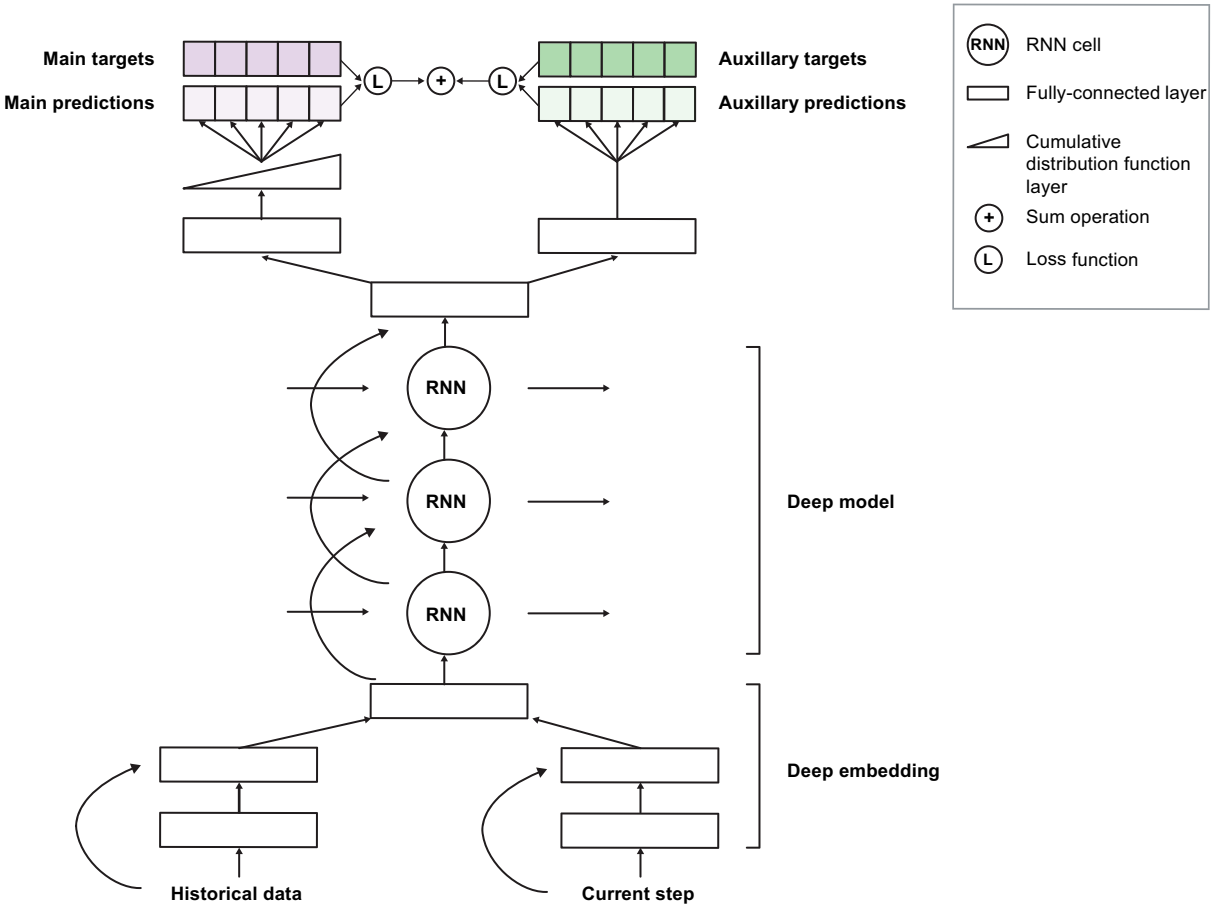
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1390-1.

Correspondence and requests for materials should be addressed to N.T. or J.R.L.

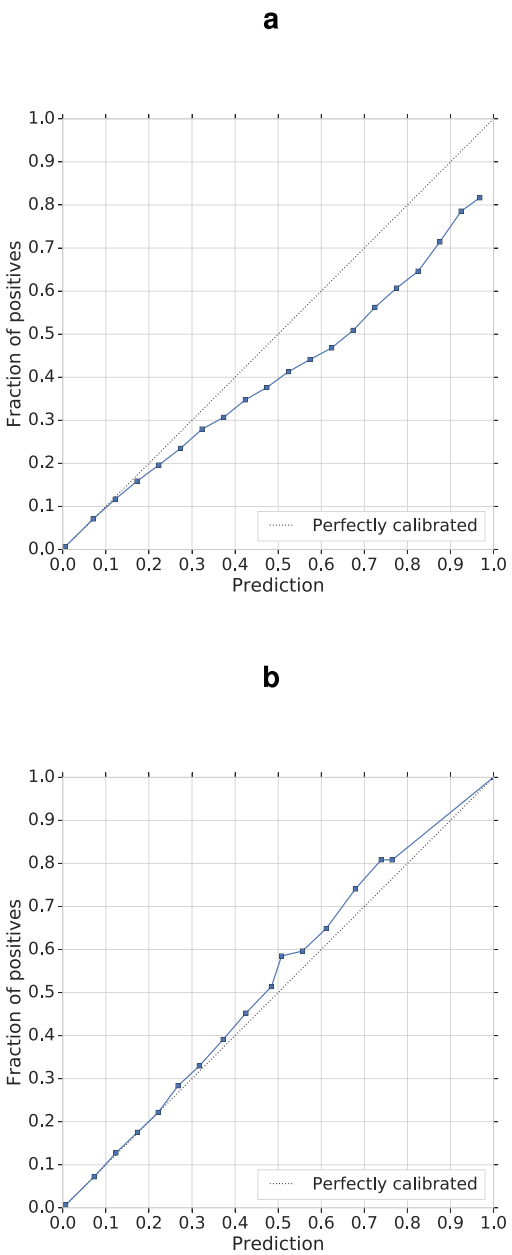
Peer review information *Nature* thanks Lui G. Forni, Suchi Saria and Eric Topol for their contribution to the peer review of this work.



Extended Data Fig. 1 | Sequential representation of electronic health record data. All electronic health record data available for each patient were structured into a sequential history for both inpatient and outpatient events in six-hourly blocks, shown here as circles. In each 24-h period, events without a recorded time were included in a fifth block. Apart from the data present at the current time step, the models optionally receive an embedding of the previous 48 h and the longer history of 6 months or 5 years.

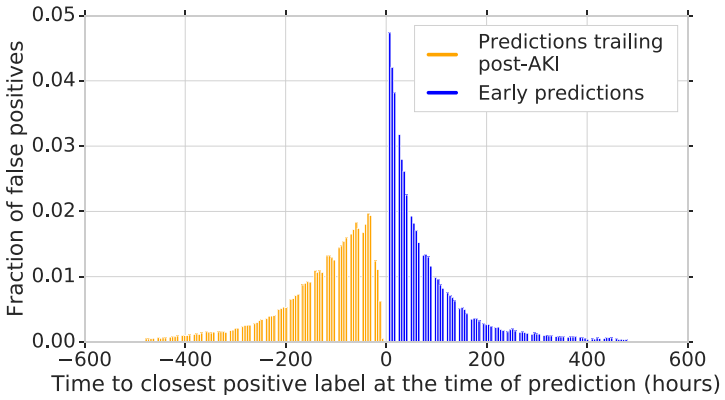


Extended Data Fig. 2 | Architecture of the proposed model. The best performance was achieved by a multi-task deep recurrent highway network architecture on top of an L_1-regularized deep residual embedding component that learns the best data representation end-to-end without pre-training.



Extended Data Fig. 3 | Calibration. a, **b**, The predictions were recalibrated using isotonic regression before (**a**) and after (**b**) calibration. Model predictions were grouped into 20 buckets, with a mean model risk prediction plotted against the percentage of positive labels in that bucket. The diagonal line demonstrates the ideal calibration.

a



b

Reason	Percent of all false positive alerts
Patients who experience AKI during admission in which the model alerts	
Model alerts >48 hours before AKI event	25%
Model alerts after AKI event	24%
Patients who do not experience AKI during admission in which model alerts	
Known renal pathology	28%
EHR evidence of clinical risk	17%
No clear risk factors from EHR	6%
Total	100%

Extended Data Fig. 4 | Analysis of false-positive predictions. a, For prediction of any AKI within 48 h at 33% precision, nearly half of all predictions are trailing, after the AKI has already occurred (orange bars) or early, more than 48 h prior (blue bars). The histogram shows the distribution of these trailing and early false positives for prediction. Incorrect predictions are mapped to their closest preceding or following episode of AKI (whichever is closer) if that episode occurs in an admission. For ±1 day, 15.2% of false positives correspond to observed AKI events within 1 day after the prediction (model reacted too early) and 2.9% correspond to observed AKI events within 1 day before the prediction (model reacted too late). **b**, Subgroup analysis for all false-positive alerts. In addition to the 49% of false-positive alerts that were made in admissions during which there was at least one episode of AKI, many of the remaining false-positive alerts were made in patients who had evidence of clinical risk factors present in their available electronic health record data. These risk factors are shown here for the proposed model that predicts any stage of AKI occurring within the next 48 h.

Extended Data Table 1 | Model performance for predicting AKI within the full range of possible prediction windows from 6 to 72 h

a

	ROC AUC [95% CI]		
Time windows	Any AKI	AKI stages 2 and 3	AKI stage 3
24h	93.4% [93.3, 93.6]	97.1% [96.9, 97.3]	98.8% [98.7, 98.9]
48h	92.1% [91.9, 92.3]	95.7% [95.5, 96.0]	98.0% [97.8, 98.2]
72h	91.4% [91.1, 91.6]	94.7% [94.4, 95.0]	97.3% [97.2, 97.6]

b

	PR AUC [95% CI]		
Time windows	Any AKI	AKI stages 2 and 3	AKI stage 3
24h	25.9% [24.6, 27.0]	36.8% [35.1, 38.7]	47.6% [45.1, 49.7]
48h	29.7% [28.5, 30.8]	37.8% [36.1, 39.6]	48.7% [46.4, 51.1]
72h	31.7% [30.6, 32.8]	37.4% [35.6, 39.1]	48.0% [46.1, 49.9]

a, **b**, With shorter time windows, closer to the actual onset of AKI, the model achieves a higher area under the receiver operating characteristic curve (ROC AUC; **a**) but a lower area under the precision–recall curve (PR AUC; **b**). This stems from different numbers of positive steps within windows of different length. These differences affect both the model precision and the false-positive rate. When making predictions across shorter time windows there is more uncertainty in the exact time of the AKI onset owing to minor physiological fluctuations, and this results in a lower precision being needed to achieve high sensitivity. Bootstrap pivotal 95% confidence intervals are calculated using $n = 200$ bootstrap samples.

Extended Data Table 2 | Daily frequency of true- and false-positive alerts when predicting different stages of AKI

a

Alert type	Frequency predicting any stage of AKI
True positive alerts	0.85% ± 0.71
False positive alerts	1.89% ± 1.20
No alerts	97.26% ± 1.63

b

Alert type	Frequency predicting KDIGO AKI stages 2 and above
True positive alerts	0.30% ± 0.35
False positive alerts	0.64% ± 0.55
No alerts	99.06% ± 0.75

c

Alert type	Frequency predicting KDIGO AKI stage 3
True positive alerts	0.27% ± 0.33
False positive alerts	0.56% ± 0.85
No alerts	99.17% ± 0.96

The frequency and standard deviation of alerts are shown for a time window of 48 h and an operating point corresponding to a 1:2 true-positive:false-positive ratio (n = 5,101 days). On an average day, clinicians would receive true-positive alerts of AKI predicted to occur within a window of 48 h ahead in 0.85% of all in-hospital patients and a false-positive prediction of a future AKI in 1.89% of patients, when predicting future AKI of any severity. Assuming none of the false positives can be filtered out and immediately discarded, clinicians would need to attend to approximately 2.7% of all in-hospital patients. For the most-severe stages of AKI, on an average day the model provides positive alerts for 0.8% of all patients. Of these, 0.27% are true positives and 0.56% are false positives. Note that there are multiple time steps at which the predictions are made within each day, so the true-positive:false-positive ratio of the daily alerts differs slightly from the stepwise ratio. **a**, Daily frequency of true and false positive alerts when predicting any stage of AKI. **b**, Daily frequency of true- and false-positive alerts when predicting KDIGO AKI stages 2 and above. **c**, Daily frequency of true- and false-positive alerts when predicting the most-severe stage of AKI (KDIGO AKI stage 3).

Extended Data Table 3 | Model performance on patients who required subsequent dialysis

Subgroup name	Sensitivity (AKI episode)	PR AUC	ROC AUC	Sensitivity (step)	Specificity (step)
In-hospital/outpatient dialysis within 30 days	84.3%	70.5%	83.5%	67.7%	83.3%
Outpatient dialysis within 90 days	90.2%	71.9%	83.8%	76.5%	76.3%

Model performance only in cases of AKI in which in-hospital or outpatient administration of dialysis was required within 30 days of the onset of AKI, or regular outpatient administration of dialysis was scheduled within 90 days. The model successfully provides early predictions for a large proportion of these cases of AKI—84.3% for cases of AKI in which there is any dialysis administration that occurs within 30 days, and 90.2% for cases in which regular outpatient administration of dialysis occurs within 90 days.

Extended Data Table 4 | Operating points for predicting AKI up to 48 h ahead of time

a

Operating points for predicting any AKI up to 48 hours ahead of time				
Precision	True positive / False positive	Sensitivity [95% CI] (AKI episode)	Sensitivity [95% CI] (step)	Specificity [95% CI] (step)
20.0%	1:4	76.7% [75.6, 77.8]	58.3% [56.9, 59.8]	94.8% [94.6, 95.1]
25.0%	1:3	68.2% [66.9, 69.7]	47.7% [46.1, 49.4]	96.8% [96.6, 97.0]
33.0%	1:2	55.8% [53.9, 57.7]	35.0% [33.3, 36.7]	98.4% [98.3, 98.5]
40.0%	2:3	46.6% [44.5, 49.0]	27.1% [25.2, 28.9]	99.1% [99.0, 99.2]
50.0%	1:1	34.7% [32.0, 37.2]	18.5% [16.7, 20.3]	99.6% [99.5, 99.6]
60.0%	3:2	24.7% [21.8, 27.3]	12.4% [10.5, 13.9]	99.8% [99.8, 99.8]
75.0%	3:1	12.0% [9.3, 14.6]	5.5% [3.9, 7.0]	100.0% [99.9, 100.0]

b

Operating points for predicting AKI stages 2 and 3 up to 48 hours ahead of time				
Precision	True positive / False positive	Sensitivity [95% CI] (AKI episode)	Sensitivity [95% CI] (step)	Specificity [95% CI] (step)
20.0%	1:4	82.0% [80.6, 83.5]	65.8% [64.0, 67.9]	98.5% [98.4, 98.6]
25.0%	1:3	77.8% [76.3, 79.7]	60.4% [58.3, 62.8]	99.0% [98.9, 99.1]
33.0%	1:2	71.4% [69.6, 73.7]	51.8% [49.6, 54.8]	99.4% [99.4, 99.5]
40.0%	2:3	65.2% [63.0, 67.7]	44.6% [42.1, 47.3]	99.6% [99.6, 99.7]
50.0%	1:1	56.2% [54.0, 59.2]	35.8% [33.5, 38.9]	99.8% [99.8, 99.8]
60.0%	3:2	45.1% [42.2, 48.6]	26.3% [23.8, 29.4]	99.9% [99.9, 99.9]
75.0%	3:1	27.5% [24.2, 31.5]	13.8% [11.7, 16.3]	100.0% [100.0, 100.0]

c

Operating points for predicting AKI stage 3 up to 48 hours ahead of time				
Precision	True positive / False positive	Sensitivity [95% CI] (AKI episode)	Sensitivity [95% CI] (step)	Specificity [95% CI] (step)
20.0%	1:4	91.2% [90.4, 92.3]	80.3% [78.4, 82.4]	98.8% [98.7, 98.9]
25.0%	1:3	88.8% [87.7, 90.1]	75.8% [73.7, 78.3]	99.1% [99.0, 99.2]
33.0%	1:2	84.1% [82.4, 85.9]	68.3% [65.7, 71.0]	99.5% [99.4, 99.5]
40.0%	2:3	79.5% [77.4, 81.8]	61.1% [57.9, 64.5]	99.7% [99.6, 99.7]
50.0%	1:1	71.3% [68.3, 74.4]	50.2% [46.4, 53.8]	99.8% [99.8, 99.8]
60.0%	3:2	61.2% [57.6, 64.9]	39.9% [35.7, 43.8]	99.9% [99.9, 99.9]
75.0%	3:1	40.5% [36.5, 46.1]	23.2% [19.6, 27.2]	100.0% [100.0, 100.0]

a, For prediction of any AKI, the model correctly provides early identification in 55.8% of all AKI episodes (when allowing for 2 false positives for every true positive) and in 34.7% of episodes if allowing for 1 false positive for every true positive. For more-severe AKI stages, it is possible to achieve a higher sensitivity for any fixed level of precision. **b**, **c**, Performance increases for prediction of AKI stages 2 and 3 (**b**), and AKI stage 3 (**c**) alone. Bootstrap pivotal 95% confidence intervals are calculated using $n = 200$ bootstrap samples for all tables.

Extended Data Table 5 | Future and cross-site generalizability experiments

a

	Patient cohorts			
Metric [95% CI]	Before t_P (test)	New admissions after t_P (test)	Subsequent admissions after t_P	All patients after t_P
Sensitivity (AKI episode)	55.09 [54.01, 56.06]	59 [57.11, 60.71]	59.04 [58.38, 59.63]	58.97 [58.33, 59.52]
ROC AUC	92.25 [92.01, 92.42]	90.19 [89.76, 90.77]	89.98 [89.83, 90.17]	89.98 [89.81, 90.14]
PR AUC	29.97 [28.61, 31.15]	30.75 [28.65, 32.81]	31.54 [30.87, 32.30]	31.28 [30.44, 32.02]
Sensitivity (step)	34.26 [33.17, 35.28]	36.87 [35.2, 38.85]	37.23 [36.67, 37.88]	37.08 [36.40, 37.65]
Specificity (step)	98.55 [98.50, 98.60]	97.66 [97.54, 97.76]	97.63 [97.58, 97.68]	97.64 [97.59, 97.68]
Precision	32.51 [31.44, 33.21]	32.66 [31.2, 34.03]	32.97 [32.52, 33.47]	32.84 [32.28, 33.33]

b

	Before t_P	After t_P
Patients		
Number of patients	599,871	246,406
Average age*	61.3	64.2
Admissions within a given period		
Unique admissions	2,134,544	364,778
ICU admissions	226,585 (10.62%)	40,102 (10.99%)
Medical admissions	1,040,923 (48.77%)	170,383 (46.71%)
Surgical admissions	373,823 (17.51%)	67,617 (18.54%)
No creatinine measured	458,486 (21.48%)	52,115 (14.29%)
Any Chronic Kidney Disease	774,883 (36.30%)	156,181 (42.82%)
Any AKI present	282,398 (13.23%)	41,950 (14.59%)

c

Metric [95% CI]	Site group *A*	Site group *B*
Sensitivity (AKI episode)	55.6% [54.5, 56.6]	54.6% [52.8, 56.3]
ROC AUC	91.8% [91.6, 92.1]	91.3% [90.8, 91.7]
PR AUC	30.0% [28.6, 31.2]	30.6% [28.3, 32.7]
Sensitivity (step)	34.3% [33.1, 35.2]	34.7% [32.6, 36.2]
Specificity (step)	98.5% [98.4, 98.5]	98.3% [98.2, 98.4]

a, Model performance when trained before the time point t_P and tested after t_P, on both the entirety of the future patient population and on subgroups of patients for whom the model has or has not seen historical information during training. The model maintains a comparable level of performance on unseen future data, with a higher level of sensitivity of 59% for a time window of 48 h ahead of the AKI episode and a precision of 2 false positives per step for each true positive. The ranges correspond to bootstrap pivotal 95% confidence intervals with $n = 200$ bootstrap samples. Note that this experiment is not a replacement for a prospective evaluation of the model. **b**, Cohort statistics for **a**, shown for before and after the temporal split t_P that was used to simulate model performance on future data. **c**, Comparison of model performance when applied to data from previously unseen hospital sites. Data were split across sites such that 80% of the data were in group A and 20% of the data were in group B. No site from group B was present in group A, and vice versa. The data were split into training, validation, calibration and test sets in the same way as in the other experiments. The table reports model performance when trained on site group A when evaluating on the test set within site group A versus the test set within site group B for predicting all AKI severities up to 48 h ahead of time. A comparable performance is seen across all key metrics. Bootstrap pivotal 95% confidence intervals are calculated using $n = 200$ bootstrap samples. Note that the model needs to be retrained to generalize from the population represented by the US Department of Veterans Affairs dataset to different demographics and sets of clinical pathways and hospital processes.

Extended Data Table 6 | Summary statistics for the data

		Training	Validation	Calibration	Test
Patients					
Unique patients		562,507	35,277	35,317	70,681
Average age*		62.4	62.5	62.4	62.3
Ethnicity	Black	106,299 (18.9%)	6,544 (18.6%)	6,675 (18.6%)	13,183 (18.7%)
	Other	456,208 (81.1%)	28,733 (81.4%)	28,642 (81.4%)	57,498 (81.3%)
Gender	Female	35,855 (6.4%)	2,300 (6.5%)	2,252 (6.4%)	4,519 (6.4%)
	Male	526,652 (93.6%)	32,977 (93.5%)	33,065 (93.6%)	66,162 (93.6%)
Diabetes		56,958 (10.1%)	3,599 (10.2%)	3,702 (10.5%)	7,093 (10.0%)
Admissions within a five year period					
Data center sites		130***	130***	130***	130***
Unique admissions		2,004,217	124,255	125,928	252,492
- per patient	Average	3.6	3.5	3.6	3.6
	Median	2	2	2	2
Duration (days)	Average	9.6	9.6	9.6	9.6
	Median	3.2	3.2	3.2	3.2
ICU admissions		214,644 (10.7%)	13,161 (10.6%)	13,411 (10.6%)	26,739 (10.6%)
Medical admissions		971,527 (48.5%)	60,762 (48.9%)	61,281 (48.7%)	121,675 (48.2%)
Surgical admissions		354,008 (17.7%)	21,857 (17.6%)	22,093 (17.5%)	44,766 (17.7%)
Renal replacement therapy		22,284 (1.1%)	1,367 (1.1%)	1,384 (1.1%)	2,784 (1.1%)
No creatinine measured		408,927 (20.4%)	25,162 (20.3%)	25,503 (20.3%)	51,484 (20.4%)
Chronic Kidney Disease	Any	746,692 (37.3%)	46,677 (37.5%)	46,622 (37.0%)	94,105 (37.3%)
	Stage 1**	8,409 (0.4%)	515 (0.4%)	576 (0.5%)	1,103 (0.4%)
	Stage 2	429,990 (21.5%)	27,162 (21.9%)	26,927 (21.4%)	54,476 (21.6%)
	Stage 3A	156,720 (7.8%)	9,837 (7.9%)	9,803 (7.8%)	19,548 (7.7%)
	Stage 3B	77,801 (3.9%)	4,675 (3.8%)	4,823 (3.7%)	9,760 (3.9%)
	Stage 4	50,535 (2.5%)	3,004 (2.5%)	3,066 (2.5%)	6,223 (2.5%)
	Stage 5	31,646 (1.6%)	1,999 (1.6%)	2,003 (1.6%)	4,098 (1.6%)
AKI present	Any AKI	267,396 (13.3%)	16,671 (13.4%)	16,760 (13.3%)	33,759 (13.4%)
	Stage 1	207,441 (10.4%)	12,794 (10.3%)	12,951 (10.3%)	26,215 (10.4%)
	Stage 2	43,446 (2.2%)	2,780 (2.2%)	2,783 (2.2%)	5,575 (2.2%)
	Stage 3	66,734 (3.3%)	4,267 (3.4%)	4,162 (3.3%)	8,453 (3.3%)

A breakdown of training (80%), validation (5%), calibration (5%) and test (10%) datasets by both unique patients and individual admissions. Where appropriate, the percentage of total dataset size is reported in parentheses. The dataset was representative of the overall US Department of Veterans Affairs population for clinically relevant demographics and diagnostic groups associated with renal pathology.

*Average age after taking into account exclusion criteria and statistical noise added to meet HIPAA Safe Harbour criteria.

**Chronic kidney disease stage 1 is evidence of renal parenchymal damage with a normal glomerular filtration rate. This is rarely recorded in our dataset; instead, the numbers for stage-1 chronic kidney disease have been estimated from admissions that carried an ICD-9 code for chronic kidney disease, but for which the glomerular filtration rate was normal. For this reason, these numbers may underrepresent the true prevalence in the population.

***In total, 172 US Department of Veterans Affairs inpatient sites and 1,062 outpatient sites were eligible for inclusion. In addition, 130 data centres aggregated data from 1 or more of these facilities, of which 114 such data centres had data for inpatient admissions used in this study. Although the exact number of sites included was not provided in the dataset for this work, no patients were excluded on the basis of location.

nature research

Corresponding author(s): Dr J R Ledsam

Last updated by author(s): June 3 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection was performed by independent members of the VA National Data Center without involvement from research team members. Collection was performed using the Vista EHR system and associated databases.
Data analysis	The networks used the TensorFlow library with custom extensions. Analysis was performed with custom code written in Python 2.7. Please see the manuscript methods section for more detail.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The clinical data used for the training, validation and test sets were collected at the US Department of Veterans Affairs and transferred to a secure data centre with strict access controls in de-identified format. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The de-identified dataset, or a test subset, may be available from the US Department of Veterans Affairs subject to local and national ethical approvals.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The dataset consisted of all eligible patients during a five year period across the entire VA healthcare system in the USA. The test population was a random selection of 10% of these, totaling 70,681 individual patients and 252,492 unique admissions (please refer to methods section for more details on how test populations were selected). A sample size requirement of 179 patients would be required to detect sensitivity and specificity at 0.05 marginal error and 95% confidence. The total number of test patients exceeded this requirement by two orders of magnitude.
Data exclusions	We excluded patients below the age of 18 and above the age of 90 in accordance with HIPAA Safe Harbor criteria, and patients without any serum creatinine recorded in EHR. (See paper methods for more detail.) To protect patient privacy sites with fewer than 250 admissions during the five year time period were also excluded; four of the 1,243 health care facilities from which the VA is composed were excluded based on this criteria. All exclusion criteria were established prior to beginning the work.
Replication	All 70,681 patients in the test set were randomly selected and were not correlated in any way. The experiments can be interpreted as 70,681 replicas of the model applied to a single patient over a fifteen year period.
Randomization	The data were randomly divided into training (80% of observations), validation (5%), calibration (5%) and testing (10%) sets. All data for a single patient was assigned to exactly one of these splits. (See paper methods for more detail.)
Blinding	When assigning patients randomly to test, validation and training groups investigators were blinded to patient covariates and all features in the EHR not required to perform the research (e.g., creatinine was required to label AKI as a ground truth). Patient recruitment was conducted by independent members of the VA National Data Center; research team members were blinded to this recruitment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The data included all VA patients aged between 18 and 90 admitted for secondary care to medical or surgical services between 10/1/2011 to 9/30/2015, with laboratory data that included serum creatinine recorded in EHR and with at least one year of EHR data prior to admission data. The test set was a randomly selected 10% of all admissions included in the work. Average age was 62.3. Males represented 93.6% of the test population. Average number of inpatient admissions was 3.6; average admission duration was 9.6 days. AKI occurred in 13.4% of admissions. These figures were consistent with the population of the VA as a whole. For more information please refer to Supplementary Table 1 in the submitted supplementary material.
Recruitment	The data was recruited from the US Department of Veterans Affairs (VA). The VA is composed of 1,243 health care facilities, including 172 VA Medical Centers and 1,062 outpatient sites of care. Aggregating data from one or more of these facilities are 130 data centres, of which 114 had data for inpatient admissions used in this study. Four sites were excluded due to small numbers of patients: fewer than 250 admissions during the fifteen year time period. No other patients were excluded based on location, and no other exclusion criteria were applied. The final dataset consisted of the records for all 703,782 patients that met inclusion and exclusion criteria. For more information please refer to the submitted manuscript.

Ethics oversight	This work, and the collection of data on implied consent, received Tennessee Valley Healthcare System Institutional Review Board (IRB) approval from the US Department of Veterans Affairs. De-identification was performed in line with the Health Insurance Portability and Accountability Act (HIPAA), and validated by the US Department of Veterans Affairs Central Database and Information Governance departments. Only de-identified retrospective data was used for research, without the active involvement of patients.

Note that full information on the approval of the study protocol must also be provided in the manuscript.