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, 3], Hugh Montgomery[4], Geraint Rees[5], Chris Laing[6], Clifton R. Baker[7], Kelly Peterson[8, 9], Ruth Reeves[7], Demis Hassabis[1], Dominic King[1], Mustafa Suleyman[1], Trevor Back[1*], Christopher Nielson[7, 10*], Joseph R. Ledsam[1*], and Shakir Mohamed[1*]

[1]DeepMind, London, UK
[2]CoMPLEX, Computer Science, University College London, London, UK
[3]Present address: University College London, London, UK
[4]Institute for Human Health and Performance, University College London, London, UK
[5]Institute of Cognitive Neuroscience, University College London, London, UK
[6]University College London Hospitals, London, UK
[7]Department of Veterans Affairs, USA
[8]VA Salt Lake City Healthcare System, USA
[9]Division of Epidemiology, University of Utah, USA
[10]University of Nevada School of Medicine, USA
[*]These authors contributed equally to this work

Early prediction of deterioration could play an important role in supporting healthcare professionals as an estimated 11% of in-hospital deaths follow a failure to promptly recognise and

treat deteriorating patients [1]. To achieve this goal requires predictions of patient risk that are continuously-updated and accurate, and which are delivered at an individual level with sufficient context, and with enough time to act. Building upon recent work modelling adverse events from electronic health records (EHR) [2–18], and taking the common and potentially life-threatening condition of Acute Kidney Injury (AKI) [19] as an exemplar, we have developed a novel deep learning approach for continuous risk prediction of future AKI. The model was developed on a large, longitudinal EHR dataset covering diverse clinical environments, comprising 703,782 adult patients across 172 inpatient and 1,062 outpatient sites. Our model predicts 55.8% of all inpatient AKI episodes, and 90.2% of all AKI that requires subsequent administration of dialysis, with a lead time of up to 48 hours and a ratio of two false alerts for every true alert. In addition to predicting future AKI, our model provides confidence assessments and a list of clinical features most salient to each prediction, alongside predicted future trajectories for clinically relevant blood tests [9]. While the recognition and prompt treatment of AKI are known to be challenging, our approach may offer new opportunities to identify patients at risk within a time window that allows early treatment.

Adverse events and clinical complications are a major cause of mortality and poor patient outcomes, and substantial effort has been made to improve their recognition [19, 20]. Few predictors have found their way into routine clinical practice, either because they lack effective sensitivity and specificity, or because they report *already existing* damage [21]. One example relates to AKI, a potentially life threatening condition affecting approximately 1 in 5 US inpatient admissions [22]. Although a significant proportion of cases are thought to be preventable with early treatment [23], current AKI detection algorithms depend on changes in serum creatinine as a marker of acute decline in renal function. Elevation of serum creatinine lags significantly behind renal injury, resulting in delayed access to treatment [24]. This supports a case for preventative 'screening' type alerts, but there is no evidence that current rule based alerts improve outcomes [25, 26]. For predictive alerts to be effective they must empower clinicians to act before major clinical decline has occurred by: (i) delivering actionable insights on preventable

conditions; (ii) being personalised for specific patients; (iii) offering sufficient contextual information to inform clinical decision-making; and (iv) being generally applicable across patient populations [27].

Promising recent work on modelling adverse events from EHR [2–18] suggests that the incorporation of machine learning may enable early prediction of AKI. Existing examples of sequential AKI risk models have either not demonstrated a clinically-applicable level of predictive performance [28] or have focused on predictions across a short time horizon, leaving little time for clinical assessment and intervention [29].

Our proposed system is a recurrent neural network (RNN) that operates sequentially over the EHR, 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 hours, although our approach can be extended to other time windows or AKI severities (see Extended Data Tables 2, 3 and 4). When the predicted probability exceeds a specified operating point threshold, the prediction is considered positive. This model was trained using data curated from a multisite retrospective dataset of 703,782 adult patients from all available sites at the US Department of Veterans Affairs (VA)—the largest integrated health care system in the United States. The dataset consisted of information available from the hospital EHR in digital format. The total number of independent entries in the dataset was approximately 6 billion, including 620,000 features. Patients were randomised across training (80%), validation (5%), calibration (5%) or 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 [19]; the incidence of KDIGO AKI was 13.4% of admissions. (Detailed descriptions of the model and dataset are provided in the Methods.)

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. Demonstrating prediction uncertainty may help clinicians distinguish ambiguous cases from

predictions fully supported by the available data. Identifying an increased risk of future AKI sufficiently in advance is critical, as longer lead times may allow preventative action to be taken. This is possible even when clinicians may not be actively intervening with, or monitoring a patient.

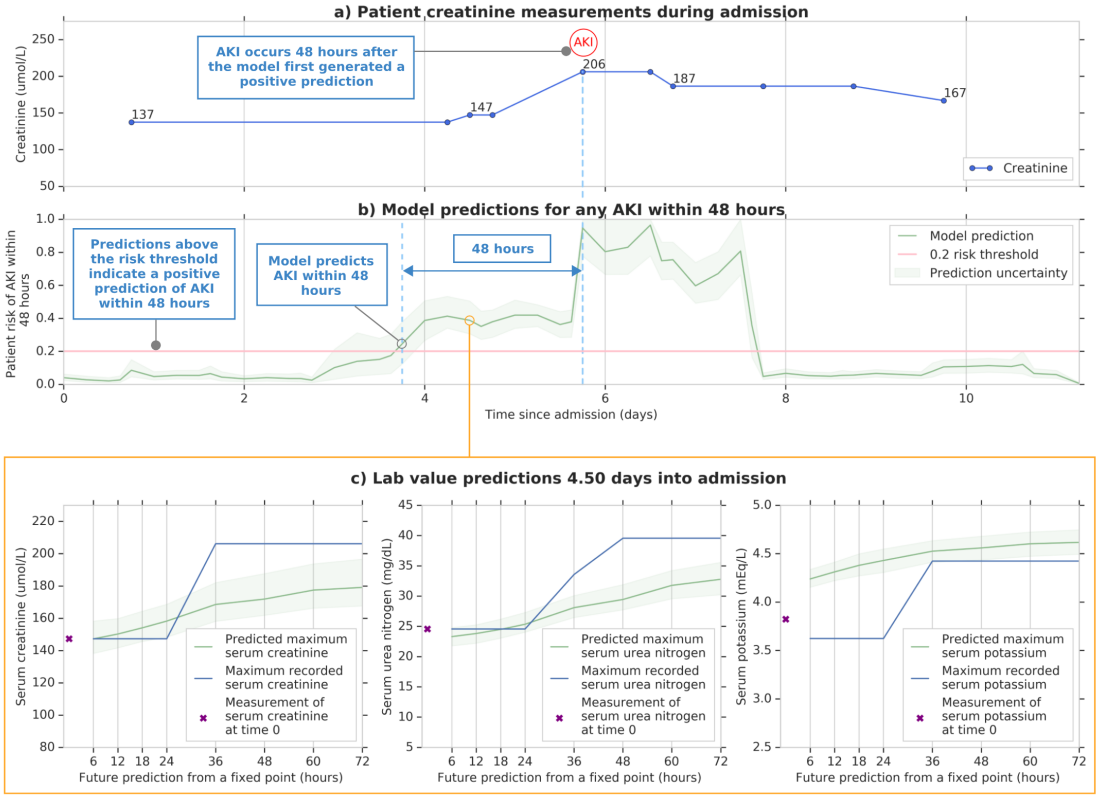


Figure 1 | Illustrative example of risk prediction, uncertainty and predicted future laboratory values. A visual representation of an 11 day admission for a 65 year old male patient with a history of chronic obstructive pulmonary disease. (**a**) The creatinine measurements throughout the admission, showing an AKI event occurring on the 5th day of admission. (**b**) The model's continuous risk predictions, where the model predicted an increase in risk of AKI onset 48 hours before it was detected according to the KDIGO criteria. A risk above 0.2, corresponding to precision of 33%, was taken as the threshold for which an AKI is predicted to occur. The lighter green borders on the risk curve indicate uncertainty, taken as the range of 100 ensemble predictions once trimmed for the highest and lowest 5 values. (**c**) Predictions made in the 4th day of admission of the maximum future observed values of serum creatinine, serum urea nitrogen, and serum potassium up to 72 hours ahead of time.

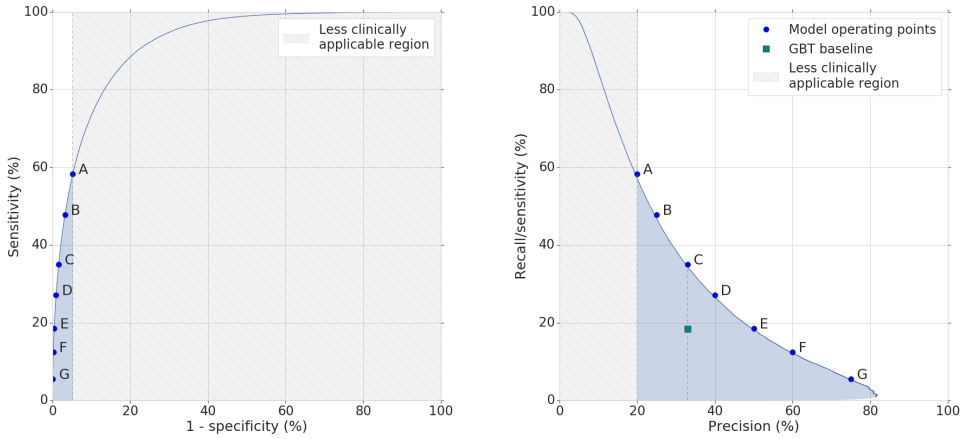
With our approach, 55.8% of inpatient AKI events of any severity were predicted early within a window of up to 48 hours in advance, with a ratio of two false predictions for every true positive. This corresponds to an area under the receiver operating characteristic curve (ROC

AUC) of 92.1% and an area under the precision-recall curve (PR AUC) of 29.7%. Set at this threshold our predictive model would, if operationalised, trigger a daily clinical assessment in 2.7% of hospitalised patients in this cohort (Extended Data 7). Sensitivity was particularly high in patients who went on to develop lasting complications as a result of AKI. The model provided early predictions correctly in 84.3% of episodes where administration of in-hospital or outpatient dialysis was required within 30 days of the onset of AKI of any stage, and 90.2% of cases where regular outpatient administration of dialysis was scheduled within 90 days of the onset of AKI (Extended Data 12). Figure 2 shows the corresponding ROC and PR curves, as well as a spectrum of different operating points of the model. An operating point can be chosen to either further increase the proportion of AKI predicted early, or reduce the percentage of false predictions at each step, according to clinical priority (Figure 3). Applied to stage 3 AKI, 84.1% of inpatient events were predicted up to 48 hours in advance, with a ratio of two false predictions for every true positive (Extended Data Table 6). To respond to these alerts on a daily basis, clinicians would need to attend to approximately 0.8% of in-hospital patients (Extended Data 7).

The model correctly identifies substantial future increases in seven auxiliary biochemical tests in 88.5% of cases (Supplementary Table 3), 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 known to be relevant to renal function (see Supplementary Table 1). The predictive performance of our model was maintained across time and hospital sites, demonstrated by additional experiments that show generalisability to data acquired at time points after the model was trained (Extended Data Tables 8, 9 and 10).

Our approach significantly outperformed ($p < 0.001$) established state-of-the-art baseline models (Supplement H). For example, a baseline model was created with gradient boosted trees (GBT) using manually-curated features known to be relevant for modelling kidney function and in routine care delivery (Supplements K and E.1), plus aggregate statistical information on trends observed in recent patient history. This yielded 3599 clinically relevant features provided to the

baselines at each step (see Methods). For the same level of precision the baseline model was able to detect 36.0% of all inpatient AKI episodes up to 48 hours ahead of time, compared to 55.8% for our model.



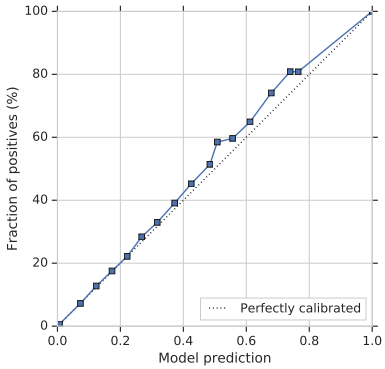
(a)

(b)

Operating points

Operating Point	Precision	TP : FP	Sensitivity (AKI episode)	Sensitivity (step)	Specificity (step)
A	20.0%	1:4	76.7% [75.6, 77.8]	58.3% [56.9, 59.8]	94.8% [94.6, 95.1]
B	25.0%	1:3	68.2% [66.9, 69.7]	47.7% [46.1, 49.4]	96.8% [96.6, 97.0]
C	33.0%	1:2	55.8% [53.9, 57.7]	35.0% [33.3, 36.7]	98.4% [98.3, 98.5]
D	40.0%	2:3	46.6% [44.5, 49.0]	27.1% [25.2, 28.9]	99.1% [99.0, 99.2]
E	50.0%	1:1	34.7% [32.0, 37.2]	18.5% [16.7, 20.3]	99.6% [99.5, 99.6]
F	60.0%	3:2	24.7% [21.8, 27.3]	12.4% [10.5, 13.9]	99.8% [99.8, 99.8]
G	75.0%	3:1	12.0% [9.3, 14.6]	5.5% [3.9, 7.0]	100.0% [99.9, 100.0]

(c)



(d)

Figure 2 | Model performance illustrated by Receiver Operating Characteristic (ROC) and Precision/Recall (PR) curves. (**a**) ROC and (**b**) PR curves for the primary outcome of predicting the risk that an AKI of any severity will occur within the next 48 hours. Blue dots correspond to operating points from (c). The grey hatched area covers the portions of ROC and PR curves that correspond to operating points with greater than four false positives for each true positive. The blue area captures the performance in the more clinically applicable part of the operating space; illustrating the higher applicability of PR Area Under Curve (AUC) for reporting model performance. The model significantly (p-value of <0.001) outperformed the gradient boosted trees baseline, shown in (b) for operating point C using two-sided Mann–Whitney U test on 200 samples per model (see Evaluation). (**c**) Different model operating points given as a fraction of AKI episodes successfully detected for different precision levels (or equivalently the TP:FP ratio) in terms of individual predictions made at each step. (**d**) Resulting calibration curve after isotonic regression for 48 hours ahead any-AKI severity predictions. Model predictions are grouped into twenty buckets, with a mean model risk prediction plotted against the percentage of positive labels in that bucket. The diagonal dotted line demonstrates the ideal calibration.

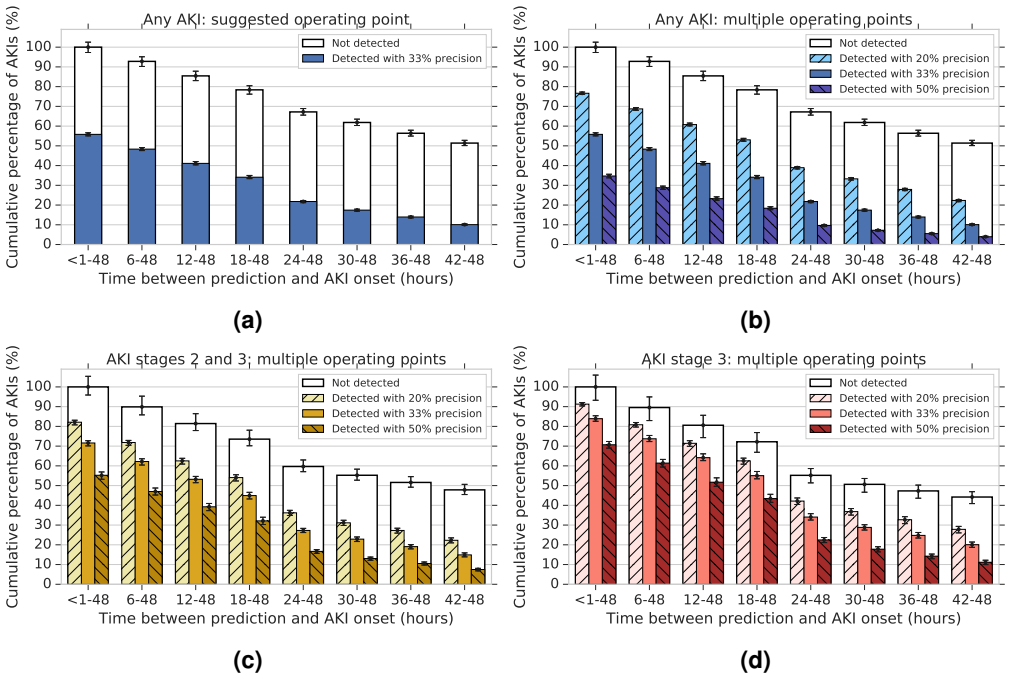


Figure 3 | The time between model prediction and actual AKI event. The models predict the risk of certain stages of AKI occurring within a particular time window. Within this window the actual time in hours between prediction and AKI event can vary. The error bars in all panels indicate 95% confidence intervals. (**a**) For the main time window studied (48 hours ahead of time) a greater proportion were correctly predicted as impending AKI events get closer to the time step immediately prior to the AKI. As AKI events often occur shortly after admission, and there is not the opportunity to predict an episode until the patient presents to hospital, the available time window in which to predict is shortened. While 100% of inpatient AKIs of each stage are possible to predict at the point of admission, fewer than 60% of all AKI events occurred more than 48 hours into an admission. (**b-d**) Model extensibility. When predicting more severe AKI stages (blue, all AKI stages; yellow, AKI stages 2 & 3; red, AKI stage 3), the model achieved higher sensitivity for the same precision. Different operating points (shown here as bars of different texture) can be configured such that more AKIs are detected early; this is demonstrated at three different precision operating points, from 20% to 50%.

Of the false positive alerts made by our model, 24.9% were positive predictions made even earlier than the 48 hour window in patients where AKI subsequently occurred (Extended Data Figure 3). 57.1% of these occurred in patients with pre-existing chronic kidney disease (CKD), 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 had already begun; such alerts can be filtered out in clinical practice. For positive risk predictions where no AKI was subsequently observed in this retrospective dataset, it is probable that many occurred in patients at risk of

AKI where appropriate preventative treatment was administered which 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 EHR that the patient required clinical review (Extended Data Table 11).

Our aim is to provide risk predictions that enable personalised preventative action to be delivered at scale. The way these predictions are used may vary by clinical setting: a trainee doctor could be alerted in real time to each patient under their care, while a specialist nephrologist or rapid response teams [30] can identify high risk patients to prioritise their response. This is possible because performance was consistent across multiple clinically important groups, notably those at an elevated risk of AKI (Supplementary Table 4). Our model is designed to complement existing routine care, as it is trained specifically to predict AKI that happened in this retrospective dataset despite existing best practices.

Although we demonstrate a model trained and evaluated on a clinically representative set of patients from the entire VA health care system, the 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 1). 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 [31] and overcome the impact of potential confounding factors related to hospital processes [32]. KDIGO is an indicator of AKI that lags long after the initial renal impairment, and model performance could be enhanced by improvements in ground-truth AKI definition and data quality. [33].

Despite state-of-the-art retrospective performance compared to existing literature, to establish clinical utility and effect on patient outcomes future work should now prospectively evaluate and independently validate the proposed model, alongside exploring its role in research into new strategies towards 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 hours 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, prior to 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 EHR datasets, we now shift to a regime where the role for machine learning in clinical care can grow rapidly, supplying new tools to enhance the patient and clinician experience, and potentially becoming a ubiquitous and integral part of routine clinical pathways.

Acknowledgements

We thank the Veterans and their families under the care of the US Department of Veterans Affairs. We would also like to thank A. Graves, O. Vinyals, K. Kavukcuoglu, S. Chiappa, T. Lillicrap, R. Raine, P. Keane, 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 VA physicians, administrators and researchers who worked on the data collection, and the rest of the DeepMind team for their support, ideas and encouragement.

G.R. & 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.L., N.T., C.N., D.H. & R.R. initiated the project.

N.T., X.G., H.A., J.L., C.N., C.B. & K.P. created the dataset.

N.T., X.G., A.S., H.A., J.R., M.Z., A.M., I.P. & S.M. contributed to software engineering.

N.T., X.G., A.M., J.R., M.Z., A.S., S.M., X.G., J.L., C.N. & C.B. analysed the results.

N.T., X.G., A.M., J.R., M.Z., S.R. & S.M. designed the model architectures.

J.L., G.R., H.M., C.L., A.C., A.K., C.H., D.K. & C.N. contributed clinical expertise.

C.M., J.L., T.B., S.M. & C.N. managed the project.

N.T., J.L., J.R., M.Z., A.M., H.M., C.B., S.M. & G.R. wrote the paper.

Competing financial interests

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

Methods

Data Description

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

The Veterans Affairs (VA) serves a population of over nine million veterans and their families across the entire United States of America. The VA is composed of 1,243 health care facilities (sites), including 172 VA Medical Centers and 1,062 outpatient facilities [34]. Data from these sites is aggregated into 130 data centres, of which 114 had data of inpatient admissions that we used in this study. Four sites were excluded since they had fewer than 250 admissions during the five year time period. No other patients were excluded based on location.

The data comprised all patients aged between 18 and 90 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 where there was at least one year of EHR data prior to admission. The data included medical records with entries up to 10 years prior to each admission date and up to two years afterwards, where available. Where available in the VA 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 one occurrence in the VA database), were excluded to ensure all potential privacy concerns were addressed. In addition, conditions that were considered sensitive were excluded prior to transfer, such as patients with HIV/AIDS, sexually transmitted diseases, substance abuse, and those admitted to mental health services.

Following this set of inclusion criteria, the final dataset comprised 703,782 patients, providing

$6,352,945,637$ clinical event entries. Each clinical entry denoted a singular procedure, laboratory test result, prescription, diagnosis etc, with $3,958,637,494$ coming from outpatient events and the remaining $2,394,308,143$ events from admissions. Extended Data Table 1 contains an overview of patient demographics in the data as well as prevalence of conditions associated with acute kidney injury 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 was assigned to exactly one of these splits.

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 6 hour period, i.e. each day was broken into four 6 hour periods and all records occurring within the same 6 hour period were grouped together. The available data within these 6 hour windows, along with additional summary statistics and augmentations, formed a feature set that formed the 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, since explicit missing value imputation in EHR predictive models does not always provide consistent improvements [35]. 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 providing this encoding of the numerical data allowed our models to process these measurements even in their absence. Discrete features like diagnostics or procedural codes were also encoded

as binary presence features.

All numerical features were normalised to the $[0, 1]$ range after capping the extreme values at the 1st and 99th percentile. This prevents the normalisation 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, lab test, vital sign, admission, transfer etc. A total of 29 such high-level concepts were present in the data. At each step, a histogram of frequencies of these concepts among the clinical entries that take 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 which 6 hour period in the day the data is associated with, were provided as explicit features to the models. In addition, we provided some simple features that make 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 hours creatinine baseline as additional numerical features. These are the baseline values that are used in the KDIGO criteria and help give important context to the models on how to interpret new serum creatinine measurements as they become available.

We additionally computed three historical aggregate feature representations at each step: one for the past 48 hours, 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 they 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 like the difference between the last observed value and the minimum/maximum and the average difference between subsequent steps that measures the temporal short-term variability of the measurement.

Because patient measurements are made irregularly, not all 6-hour time periods in a day will have new data associated with them. Our models still operate at regular time intervals, 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 utilise the patterns of missingness in the data during the training process.

For about 35% of all entries, the day on which they occurred was known, but not the specific time during the day. 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 utilise 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 EHR prior to 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, where they can be factored in for future admissions. This discards potentially useful information, so 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 based on the KDIGO [19] criteria, the recommendations of which are based on systematic reviews of relevant trials. KDIGO accepts three definitions of AKI: an increase in serum creatinine of 0.3mg/dl (26.5μmol/l) within 48 hours; an increase in serum creatinine of 1.5 times a patient's baseline creatinine level, known or presumed to have occurred within the prior 7 days; or a urine output of <0.5 ml/kg/h over 6 hours [19]. The first two definitions were used to provide ground truth labels for the onset of an 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 annualised creatinine was used where previous measurements where available; where these were not present the Modification of Diet in Renal Disease (MDRD) formula was applied to estimate a 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 & 3), 'moderate and severe AKI' (KDIGO stages 2 & 3), and 'severe AKI' (KDIGO stage 3).

The AKI stages were computed at times when 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. In order to avoid basing the current estimate of the KDIGO AKI stage on an old measurement that may no longer be reliable, the AKI states were propagated for at most 4 days forward in case no new creatinine measurements were observed. From that point onwards they were marked as unknown. Patients experiencing acute kidney injury 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 quite uncommon. A gap of 4 days between subsequent creatinine measurements represents the 95th percentile in the distribution of time between two consecutive creatinine measurements.

The prediction target at each point in time is a binary variable that is positive if the AKI category of interest (e.g., 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 use eight future time horizons, $6\text{h}, 12\text{h}, 18\text{h}, 24\text{h}, 36\text{h}, 48\text{h}, 60\text{h},$ and 72h ahead, which are all available at each time point.

Event sequences of patients undergoing renal replacement therapy (RRT) were excluded from the target labels heuristically based on the data entries of RRT procedures being performed in the EHR, for the duration of dialysis administration. We have excluded entire subsequences of events between RRT procedure entries that occur within a week of each other. The edges of the subsequence were also appropriately excluded from label computations.

Predictive Models of AKI

Our predictive system operates sequentially over the electronic health record. At each time point, input features, which we described above, were provided to a statistical model whose output is a probability of any-severity stage of AKI occurring in the next 48 hours. 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 the next section. The novelty 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 EHR 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, i.e. the parameters can be learned jointly without pretraining any parts of the model. To provide useful predictions, we train an ensemble of predictors to estimate the model's confidence, and the resulting ensemble predictions are then calibrated using isotonic regression to reflect the frequency of observed outcomes [36].

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 (ReLU) activations. We use L_1 regularisation 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 significant advantages in using these methods.

RNN core. Recurrent neural networks (RNNs) run sequentially over the EHR 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 [37], which at each time step takes the embedding vector as an input. We use the Simple Recurrent Unit (SRU) 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 (LSTM) [38], update gate RNN (UGRNN) and Intersection RNN [39], simple recurrent units (SRU) [40, 41], gated recurrent units (GRU) [42], the Neural Turing Machine (NTM) [43], memory-augmented neural network (MANN) [44], the Differentiable Neural Computer (DNC) [45], and the Relational Memory Core (RMC) [46]. These alternatives did not provide significant performance improvements over the SRU architecture (see Supplement H).

Prediction targets and training objectives. The output of the RNN is fed to a final linear prediction layer that makes predictions over all 8 future prediction windows (6 hour windows from 6 hours ahead to 72 hours ahead). We use a cumulative distribution function layer (CDF) across different time windows to encourage monotonicity, since 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, where 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 used to make the future AKI predictions. The laboratory tests predicted are ones known to be relevant to kidney function, specifically: creatinine, urea nitrogen, sodium, potassium, chloride, calcium and phosphate. This multitask approach results in better generalisation and more robust representations, especially under class imbalance [47–49]. The overall improvement we observed from including the auxiliary task was around 3% PR AUC in most cases (see Supplementary Table 10 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, prediction for positive labels contributes proportionally more to the total loss.

Training and hyperparameters. We selected our proposed model architecture among several alternatives based on the validation set performance (see Supplement G) and have subsequently performed an ablation analysis of the design choices (see Supplement I). All variables are initialised via normalised (Xavier) initialisation [50] and trained using the Adam optimisation scheme [51]. We employ 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 backpropagation 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 Table 8. We conducted extensive hyperparameter explorations of dropout rates for different kinds of dropout to determine the best model regularisation. We have considered input dropout, output dropout, embedding dropout, cell state dropout and variational dropout. None of these had 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-regularised logistic regression or gradient boosted trees (GBTs), trained on a clinically relevant set of features 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 (see Supplement E.1) 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 directly associated with an increased risk of AKI. The full feature set is listed in Supplement K. We additionally computed a set of manually-engineered features (creatinine yearly and 48-hourly baselines in line with KDIGO guidelines, ratio of blood urea nitrogen to serum creatinine, grouped severely reduced GFR corresponding to stages 3a - 5, flagging diabetic patients by combining ICD9 codes and values of measured haemoglobin A1c) and a representation of the patient's short-term and long-term history (see Section Feature Representation). These features were provided explicitly, since the interaction terms and historical trends might not have been recovered by simpler models. This resulted in a total of 3599 possible features for the baseline model. We provide a table with a full set of baseline comparison in supplement H.

Evaluation

The data was split into training, validation, calibration and test sets in such a way that information from a given patient is 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/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 [52]. For calibration we considered Platt scaling [53] and Isotonic Regression [36]. To compare uncalibrated predictions to recalibrated ones we used the Brier score [54] and reliability plots [55]. The

best models were finally evaluated on the independent test set that was held out 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 (PR AUC), the area under the receiver-operating curve (ROC AUC), 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 hours. In contrast, the precision is computed per-step since the predictions are made at each step, to account for the rate of false alerts over time.

Due 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 different AKI episodes may offer a different number of such early alerting steps depending on how late they occur within the admission. AKIs occurring later during in-hospital stay can be predicted earlier than those that occur immediately upon admission. To better assess the clinical applicability of the proposed model we explicitly compute the AKI episode sensitivity for different levels of step-wise precision.

Given that the models were designed for continuous monitoring and risk prediction, they were evaluated at each 6-hour 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 where there was no record of AKI occurring in the relevant future time window were considered as 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 (ROC AUC) as well as the area under the precision-recall curve (PR AUC). PR AUC is known to be more informative for class-imbalanced predictive tasks [56], as it is more sensitive to changes in the number of false positive predictions.

To gauge uncertainty on a trained model's performance we calculated 95% confidence intervals with the pivot bootstrap estimator [57]. 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 Mann–Whitney U test (two-sided) [58] 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 [59] and medical imaging predictions [60]. 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-value < 0.01, see Supplement B).

Reporting Summary

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

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 was used for research, without the active involvement of patients.

Code Availability

We make use of several open-source libraries to conduct our experiments, namely the machine learning framework TensorFlow[1] along with the TensorFlow library Sonnet[2] which provides implementations of individual model components [61]. 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 section and in the supplementary figures to allow for independent replication.

Data Availability

The clinical data used for the training, validation and test sets was collected at the US Department of Veterans Affairs and transferred to a secure data centre with strict access controls in de-identified format. Data was 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.

[1] https://github.com/tensorflow/tensorflow
[2] https://github.com/deepmind/sonnet

Abbreviations

Abbreviation	Description
AE	Autoencoder
AKI	Acute Kidney Injury
AKIN	Acute Kidney Injury Network
AUC	Area Under Curve
BIDMC	Beth Israel Deaconess Medical Center
CDF	Cumulative Distribution Function
CKD	Chronic Kidney Disease
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CPT	Current Procedural Terminology
DNC	Differentiable Neural Computer
ED	Emergency Department
EHR	Electronic Health Record
ER	Emergency Room
GAM	Generalised Additive Model
GBT	Gradient Boosted Trees
GFR	Glomerular Filtration Rate
GRU	Gated Recurrent Unit
GP	Gaussian Processes
HIPAA	Health Insurance Portability and Accountability Act
ICD-9	International Statistical Classification of Diseases and Related Health Problems
ICU	Intensive Care Unit
IRB	Institutional Review Board
ITU	Intensive Treatment Unit
IV	Intravenous Therapy
KDIGO	Kidney Disease: Improving Global Outcomes guidelines
LOINC	Logical Observation Identifiers Names and Codes
LR	Logistic Regression
LSTM	Long Short-Term Memory Network
MANN	Memory-Augmented Neural Network
MDP	Markov Decision Process
MLP	Multilayer Perceptron
NHSE	National Health Service England
NPV	Negative Predictive Value
NTM	Neural Turing Machine
PPV	Positive Predictive Value
PR	Precision/Recall
ReLU	Rectified Linear Unit
RF	Random Forest
RIFLE	Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease
RNN	Recurrent Neural Network
RMC	Relational Memory Core
ROC	Receiver Operating Characteristic
RRT	Renal Replacement Therapy
SMC	Stanford Medical Centre
SRU	Simple Recurrent Unit
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
UGRNN	Update Gate Recurrent Neural Network
VA	US Department of Veterans Affairs
VAE	Variational Autoencoder
WCC	White Cell Count

Extended data legends

Extended Data Figure 1 | The sequential representation of EHR data. All EHR data available for each patient was structured into a sequential history for both inpatient and outpatient events in six hourly blocks, shown here as circles. In each 24 hour 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 hours and the longer history of 6 months or 5 years.

Extended Data Figure 2 | The proposed model architecture. The best performance was achieved by a multitask deep recurrent highway network architecture on top of an L1-regularised deep residual embedding component that learns the best data representation end-to-end without pre-training.

Extended Data Figure 3 | Early and trailing positive predictions. For the prediction of AKI within 48 hours, nearly half of all predictions are made either (**a**) after the AKI has already occurred, or (**b**) more than 48 hours prior to the AKI. The histogram shows the full distribution of these trailing and early false positive predictions, for prediction of any AKI within 48 hours at 33% precision. Incorrect predictions above the set alerting threshold are mapped to their closest preceding/following AKI episode (whichever is closer) if there is one in an admission. For $\pm$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 prior to the prediction (model reacted too late).

Extended Data Table 1 | Summary statistics for the data. A breakdown of training (80%), validation (5%), calibration (5%) and test (10%) datasets by both unique patients and individual admissions. Where appropriate, percent of total dataset size is reported in parentheses. The

dataset was representative of the overall VA population for clinically relevant demographics and diagnostic groups associated with renal pathology.

Extended Data Table 2 | Model performance for predicting any severity of AKI within the full range of possible prediction windows from 6-72 hours. On shorter time windows, closer to the actual onset of AKI, the model achieves a higher ROC AUC but lower PR AUC. This difference in the metrics stems from the different number of positive steps within the windows of different length. For longer windows, there are more time steps where AKI occurs within the time window. 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 due to minor physiological fluctuations and this results in a lower precision being needed in order to achieve high sensitivity.

Extended Data Table 3 | Model ROC AUC performance. ROC AUC performance when predicting the risk of future AKI, for all AKI severities across different time windows.

Extended Data Table 4 | Model PR AUC performance. PR AUC performance when predicting the risk of future AKI, for all AKI severities across different time windows.

Extended Data Table 5 | Example operating points for predicting AKI stages 2 and 3 up to 48 hours ahead of time. The model correctly identifies 71.4% of all AKI stage 2 or 3 episodes early if allowing for two false positives for every true positive, and 56.2% if allowing for one 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.

Extended Data Table 6 | Operating points for predicting AKI stage 3 up to 48 hours ahead of time. The model identifies 84.1% of all AKI stage 3 episodes early if allowing for

two false positives for every true positive, and 71.3% when allowing for one false positive for every true positive.

Extended Data Table 7 | Daily frequency of true and false positive alerts when predicting different stages of AKI. The frequency of alerts and its standard deviation are shown for a time window of 48 hours an operating point corresponding to a 1:2 TP:FP ratio (N=5101 days). On an average day, clinicians would receive true positive alerts of AKI predicted to occur within a window of 48 hours 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 the 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, the model alerts on an average day in 0.8% of all patients. Of those, 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 TP:FP ratio of the daily alerts differs slightly from the step-wise ratio.

Extended Data Table 8 | Generalisability to future data. Model performance when trained before the time point t_P and tested after t_P, both on the entirety of the future patient population as well as subgroups of patients for which the model has or hasn't 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 hours ahead of time and a precision of two false positives per step for each true positive. Note that this experiment is not a replacement for a prospective evaluation of the model.

Extended Data Table 9 | Cohort statistics for Extended Data Table 8. Dataset statistics are shown for both before and after the temporal split t_P that was used to simulate model performance on future data.

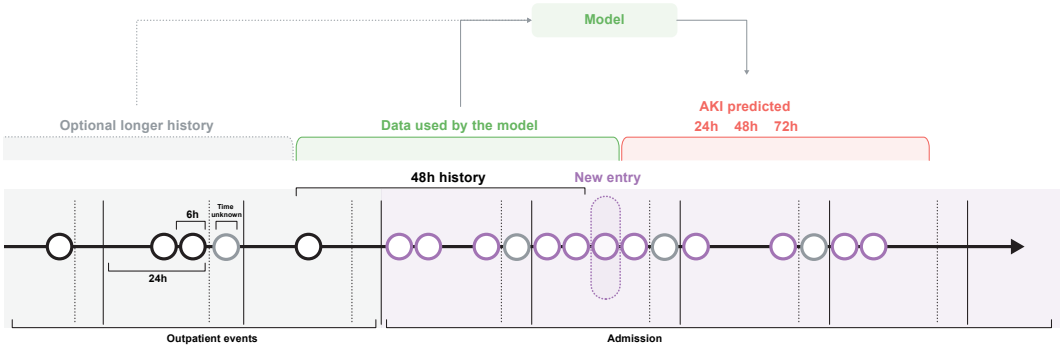
Extended Data Table 10 | Cross-site generalisability. Comparison of mode performance when applied to data from previously unseen hospital sites. Data was split across sites so that 80% of the data was in group *A* and 20% in group *B*. No site from group *B* was present in group *A* and vice versa. The data was split into training, validation, calibration and test 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 hours ahead of time. No statistically significant difference in performance was seen across most of the key metrics. Note that the model would still need to be retrained to generalise outside of the VA population to a different demographic and a different set of clinical pathways and hospital processes elsewhere.

Extended Data Table 11 | Subgroup analysis for all false positive alerts. In addition to the 49% made in admissions during which there was at least one AKI episode many of the remaining false positive alerts were made in patients with evidence of clinical risk factors present in the EHR data available. These risk factors are shown here for the proposed model predicting any stage of AKI within the next 48 hours.

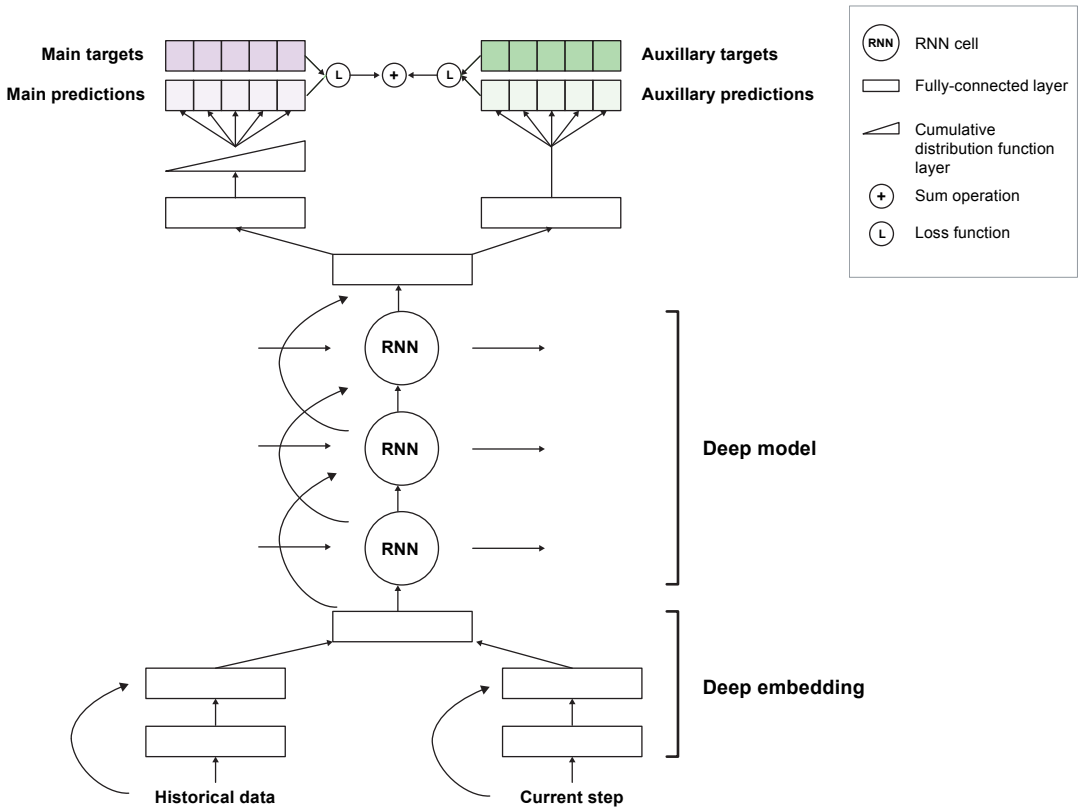
Extended Data Table 12 | Model performance on patients requiring subsequent dialysis. Model performance only in AKI cases where either in-hospital or outpatient administration of dialysis is required within 30 days of the onset of AKI, or where regular outpatient administration of dialysis is scheduled within 90 days. The model successfully predicts a large proportion of these AKI cases early, 84.3% of AKI cases where there is any dialysis administration occurring within 30 days and 90.2% of cases where regular outpatient administration of dialysis occurs within 90 days.

Extended data

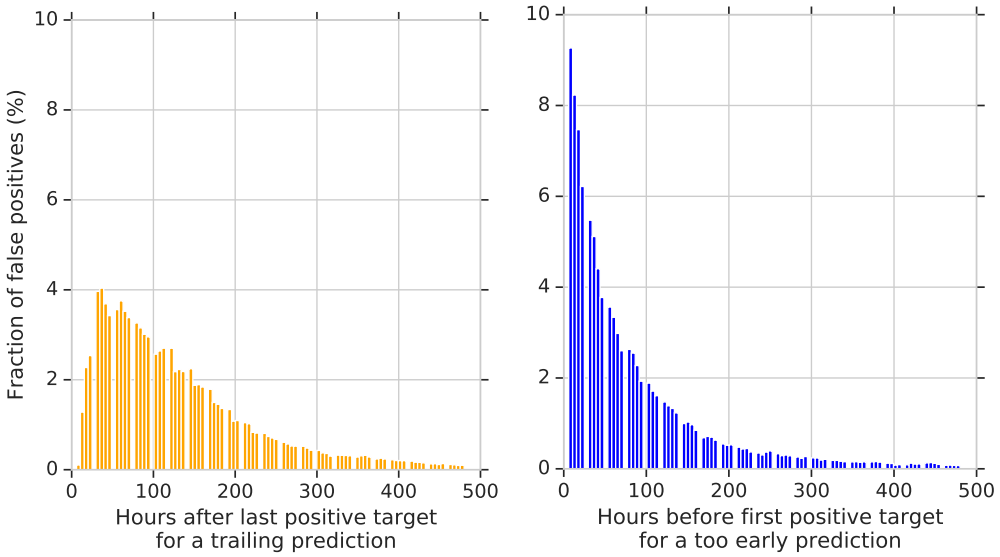
Extended data figures



Extended Data Figure 1 | The sequential representation of EHR data. All EHR data available for each patient was structured into a sequential history for both inpatient and outpatient events in six hourly blocks, shown here as circles. In each 24 hour 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 hours and the longer history of 6 months or 5 years.



Extended Data Figure 2 | The proposed model architecture. The best performance was achieved by a multitask deep recurrent highway network architecture on top of an L1-regularised deep residual embedding component that learns the best data representation end-to-end without pre-training.



Extended Data Figure 3 | Early and trailing positive predictions. For the prediction of AKI within 48 hours, nearly half of all predictions are made either (**a**) after the AKI has already occurred, or (**b**) more than 48 hours prior to the AKI. The histogram shows the full distribution of these trailing and early false positive predictions, for prediction of any AKI within 48 hours at 33% precision. Incorrect predictions above the set alerting threshold are mapped to their closest preceding/following AKI episode (whichever is closer) if there is one in an admission. For $\pm$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 prior to the prediction (model reacted too late).

Extended data tables

Extended Data Table 1 | Summary statistics for the data. A breakdown of training (80%), validation (5%), calibration (5%) and test (10%) data splits by both unique patients and individual admissions. Where appropriate, percent of total data size is reported in parentheses. The total dataset was representative of the overall VA population for clinically relevant demographics and diagnostic groups associated with renal pathology.

		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%)

*Average age after taking into account exclusion criteria and statistical noise added to meet HIPAA Safe Harbor criteria **CKD stage 1 is evidence of renal parenchymal damage with a normal glomerular filtration rate (GFR). This is rarely recorded in our dataset; instead the numbers for stage 1 CKD have been estimated from admissions that carried an ICD-9 code for CKD, but where GFR was normal. For this reason these numbers may under-represent the true prevalence in the population.

***172 VA inpatient sites and 1,062 outpatient sites were eligible for inclusion. 130 data centres aggregate data from one or more of these facilities, of which 114 such data centres had data for inpatient admissions used in this study. While the exact number of sites included was not provided in the dataset for this work, no patients were excluded based on location.

Extended Data Table 2 | Model performance for predicting any severity of AKI within the full range of possible prediction windows from 6-72 hours. On shorter time windows, closer to the actual onset of AKI, the model achieves a higher ROC AUC but lower PR AUC. This difference in the metrics stems from the different number of positive steps within the windows of different length. For longer windows, there are more time steps where AKI occurs within the time window. 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 due to minor physiological fluctuations and this results in a lower precision being needed in order to achieve high sensitivity.

Prediction window	ROC AUC [95% CI]	PR AUC [95% CI]
6 hours	95.9% [95.8, 96.0]	13.8% [13.0, 14.5]
12 hours	94.9% [94.8, 95.1]	20.5% [19.5, 21.5]
18 hours	94.1% [94.0, 94.3]	23.8% [22.7, 24.9]
24 hours	93.4% [93.3, 93.6]	25.9% [24.6, 27.0]
36 hours	92.8% [92.6, 92.9]	28.5% [27.3, 29.6]
48 hours	92.1% [91.9, 92.3]	29.7% [28.5, 30.8]
60 hours	91.7% [91.5, 91.9]	30.9% [29.8, 32.0]
72 hours	91.4% [91.1, 91.6]	31.7% [30.6, 32.8]

Extended Data Table 3 | Model ROC AUC performance. ROC AUC performance when predicting the risk of future AKI, for all AKI severities across different time windows.

	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]

Extended Data Table 4 | Model PR AUC performance. PR AUC performance when predicting the risk of future AKI, for all AKI severities across different time windows.

	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]

Extended Data Table 5 | Example operating points for predicting AKI stages 2 and 3 up to 48 hours ahead of time. The model correctly identifies 71.4% of all AKI stage 2 or 3 episodes early if allowing for two false positives for every true positive, and 56.2% if allowing for one 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.

		Operating points		
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]

Extended Data Table 6 | Operating points for predicting AKI stage 3 up to 48 hours ahead of time. The model identifies 84.1% of all AKI stage 3 episodes early if allowing for two false positives for every true positive, and 71.3% when allowing for one false positive for every true positive.

		Operating points		
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]

Extended Data Table 7 | Daily frequency of true and false positive alerts when predicting different stages of AKI. The frequency of alerts and its standard deviation are shown for a time window of 48 hours an operating point corresponding to a 1:2 TP:FP ratio (N=5101 days). On an average day, clinicians would receive true positive alerts of AKI predicted to occur within a window of 48 hours 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 the 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, the model alerts on an average day in 0.8% of all patients. Of those, 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 TP:FP ratio of the daily alerts differs slightly from the step-wise ratio.

(a) Daily frequency of true and false positive alerts when predicting any stage of AKI

Alert type	Frequency
True positive alerts	0.85% $\pm$ 0.71
False positive alerts	1.89% $\pm$ 1.20
No alerts	97.26% $\pm$ 1.63

(b) Daily frequency of true and false positive alerts when predicting KDIGO AKI stages two and above

Alert type	Frequency
True positive alerts	0.30% $\pm$ 0.35
False positive alerts	0.64% $\pm$ 0.55
No alerts	99.06% $\pm$ 0.75

(c) Daily frequency of true and false positive alerts when predicting the most severe stage of AKI - KDIGO AKI stage 3

Alert type	Frequency
True positive alerts	0.27% $\pm$ 0.33
False positive alerts	0.56% $\pm$ 0.85
No alerts	99.17% $\pm$ 0.96

Extended Data Table 8 | Generalisability to future data. Model performance when trained before the time point t_P and tested after t_P, both on the entirety of the future patient population as well as subgroups of patients for which the model has or hasn't 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 hours ahead of time and a precision of two false positives per step for each true positive. Note that this experiment is not a replacement for a prospective evaluation of the model.

	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]

Extended Data Table 9 | Cohort statistics for Extended Data Table 8. Dataset statistics are shown for both before and after the temporal split t_P that was used to simulate model performance on future data.

		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%)
Chronic Kidney Disease	Any	774,883 (36.30%)	156,181 (42.82%)
AKI present	Any AKI	282,398 (13.23%)	41,950 (14.59%)

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

Extended Data Table 10 | Cross-site generalisability. Comparison of model performance when applied to data from previously unseen hospital sites. Data was split across sites so that 80% of the data was in group *A* and 20% in group *B*. No site from group *B* was present in group *A* and vice versa. The data was split into training, validation, calibration and test 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 hours ahead of time. No statistically significant difference in performance was seen across most of the key metrics. Note that the model would still need to be retrained to generalise outside of the VA population to a different demographic and a different set of clinical pathways and hospital processes elsewhere.

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]

Extended Data Table 11 | Subgroup analysis for all false positive alerts. In addition to the 49% made in admissions during which there was at least one AKI episode many of the remaining false positive alerts were made in patients with evidence of clinical risk factors present in the EHR data available. These risk factors are shown here for the proposed model predicting any stage of AKI within the next 48 hours.

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 Table 12 | Model performance on patients requiring subsequent dialysis. Model performance only in AKI cases where either in-hospital or outpatient administration of dialysis is required within 30 days of the onset of AKI, or where regular outpatient administration of dialysis is scheduled within 90 days. The model successfully predicts a large proportion of these AKI cases early, 84.3% of AKI cases where there is any dialysis administration occurring within 30 days and 90.2% of cases where regular outpatient administration of dialysis occurs within 90 days.

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%

Further Supplementary Information

A Clinically Applicable Approach to the Continuous Prediction of Future Acute Kidney Injury

The aim of this supplementary information is to provide further information to support the claims made in the letter "*A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury*. It is the hope of the authors that by providing these supplementary results and associated discussion that the conclusions of the letter are strengthened, along with the reproducibility of the work.

In addition to the Extended Data we present the following supplementary material:

- Supplements A - C provide an analysis of the additional information provided by our proposed model to aid interpretation of the AKI predictions.
- Supplement D shows model performance across multiple clinically important groups.
- Supplement E provides and an extensive review of the literature into AKI risk models and machine learning and deep learning for electronic health records.
- Supplement F shows systematically selected case examples for both correct and incorrect model predictions.
- Supplements G-K provide additional technical information of interest to those wishing to reproduce the findings reported in not suitable for inclusion in the letter "*A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury*. These supplements are included only for editorial review, and will be removed to feature only in an accompanying protocol paper, alongside further discussion of parts of the Extended Data.

A. Feature saliency

Knowing that the predictions of future AKI risk are derived from clinical entries that can be meaningfully associated with future acute kidney injury increases confidence in the correctness of the predictive models and their robustness to potential confounders in the data.

We have investigated the significance of individual features in our trained models based on occlusion analysis [62]. Masking out individual features can lead to either an increase or a decrease in the predicted risk of future AKI. The results are shown in Supplementary Table 1. There exist other ways of looking at feature saliency and prior studies had often approached this problem by looking at the magnitudes of model parameters relating to features, or looking at the gradient of the model's risk output with respect to the input features [63]. These approaches are not well defined when comparing across both numerical and categorical features, which is why we have opted for the occlusion approach instead, as it is a more principled way of handling such data as present in our EHR feature representation at each step.

Supplementary Table 1 | The significance of individual features in our proposed model. The ten most salient features across all predictions are shown as determined by occlusion analysis. Many salient features come from laboratory tests associated with renal function, vital signs, as well as procedures associated with an increased risk of renal complications. As could be expected when predicting future AKI, changes in creatinine were the most salient amongst the frequently sampled features.

Feature name	Feature type	Correlation direction
Serum creatinine yearly baseline	numerical	negative
Serum creatinine 48h baseline	numerical	negative
Low serum calcium	presence	positive
Lab results available	aggregate count	negative
Malignant neoplasm of kidney	presence	positive
Emergency department visit	presence	negative
Procedure: rechanneling of artery	presence	positive
Serum creatinine	numerical	negative
pH (arterial blood gas)	numerical	positive
Total knee arthroplasty	presence	positive

Many salient features come from laboratory tests associated with renal function, vital signs, as well as procedures associated with an increased risk of renal complications. As could be expected when predicting future AKI, changes in creatinine were the most salient amongst the frequently sampled features. The negative correlation of an increase in values of serum crea-

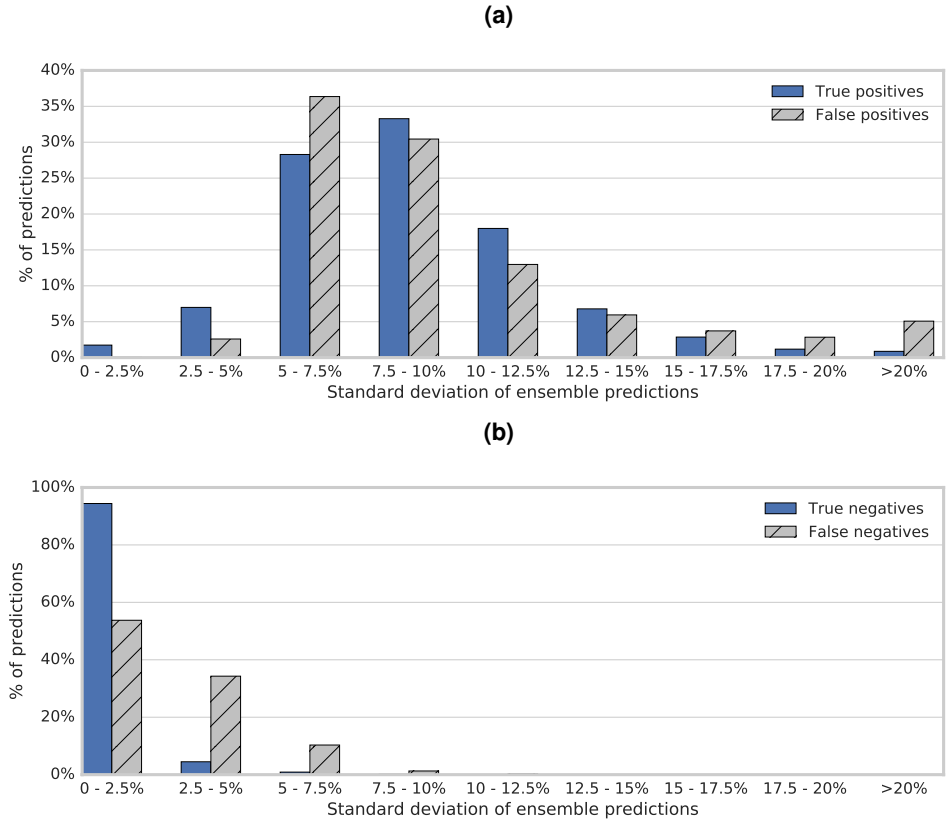
tinine baselines shown in Supplementary Table 1 is indicative of the fact that KDIGO is less likely to interpret a given increase in creatinine as an AKI if the baselines are higher, as it is based on relative increases over the baselines. Concentrations of serum calcium that are either substantially higher or lower than normal are known to be associated with kidney disease. The number of laboratory tests being taken is negatively correlated with AKI risk, which may indicate that closer patient monitoring is more likely to identify issues early and provide treatment that reduces the risk of AKI.

Higher concentrations of serum creatinine are indicative of an increased risk of future AKI in cases when the models are making positive predictions. It is therefore interesting to observe the negative average correlation reported in Supplementary Table 1. Higher baseline levels of serum creatinine may be associated with a lower risk of KDIGO AKI in patients that do not go on to develop AKI within the admission.

B. Prediction uncertainty

The ability to provide a measure of confidence in model predictions has important practical consequences. This additional information can help clinicians interpret the individual model predictions and the variance contained within them. Here we demonstrate that the predictions the model is more confident in are more likely to be correct.

Supplementary Figure 1 illustrates the relationship between model confidence and prediction accuracy. The model is generally less confident when it makes mistakes: the confidence is lower (p-value < 0.01) in false positive predictions than true positive predictions and false negative predictions than true negative predictions, as measured by the mean standard deviation of ensemble risk.



Supplementary Figure 1 | The relationship between model confidence and prediction accuracy. The two histograms demonstrate the standard deviation in predictions from an ensemble for different outcomes, shown here for an ensemble of models predicting the occurrence of an AKI of any severity within the next 48 hours. Figure **a** shows that for true positive predictions (N=67,546), the mean standard deviation (95% confidence interval: [0.880, 0.882]) is significantly lower than the mean standard deviation (95% confidence interval: [0.966, 0.968]) for false positives (N=128,292) as evidenced by a 2-sided T-test (p-value < 0.01). Figure **b** shows that for true negative predictions (N=8,907,932), the mean standard deviation (95% confidence interval: [0.005, 0.005]) is significantly lower than the mean standard deviation (95% confidence interval: [0.026, 0.026]) for false negatives (N=127,062) as evidenced by a 2-sided T-test (p-value < 0.01).

C. Performance on auxiliary tasks

In our experiment we used a set of auxiliary numerical prediction tasks along with the main task of predicting KDIGO AKI ahead of time. In particular, at each step the models were also asked to predict the maximum future observed values of seven biochemical tests of renal function for the same set of time intervals as used to make future AKI predictions. For these lab tests,

an increase in value usually signifies a worsening of kidney function, and is why predicting the maximum future values becomes relevant in understanding the evolution of kidney function over time.

Supplementary Table 2 shows the prediction performance as the relative and absolute L1 error for model predictions of the selected laboratory values 48 hours ahead of time. The mean absolute error is substantially lower than the standard deviation of the measurements for all laboratory values being predicted. The performance of the proposed recurrent neural network architecture is substantially higher than the performance of the logistic regression baseline in predicting these future lab values.

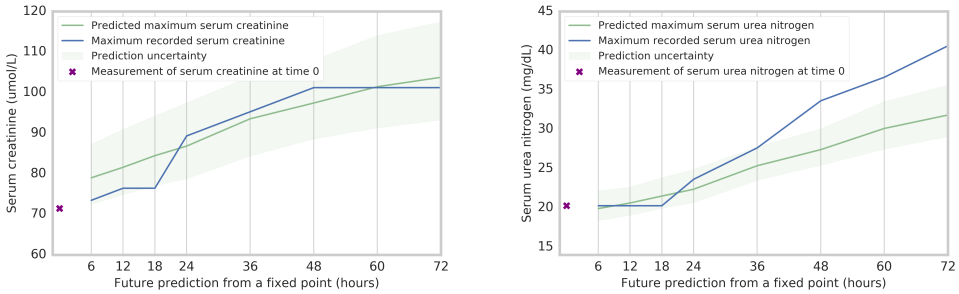
Supplementary Table 3 shows the accuracy of the model in predicting the trajectory of the selected laboratory values 48 hours ahead of time. Supplementary Figure 2 shows an example of these predictions for a given admission.

Supplementary Table 2 | Model performance for the auxiliary task of predicting the maximum future observed values of a set of seven laboratory values within 48 hours. A comparison is made between the relative prediction error for a logistic regression baseline model and a chosen recurrent neural network (SRU). Ranges indicate the 95% confidence interval.

Laboratory test	Units	Subgroup	Number samples (000s)	Subgroup mean	Subgroup standard deviation	Absolute error (SRU)	Relative error (%) (SRU)	Relative error (%) (LR)
Serum urea nitrogen	mg/dL	Population	2912.4	21.6	14.5	3.4	18.7 [18.6, 18.7]	89.7 [69.0, 101.6]
		AKI in 48 hours	188.9	36.4	19.8	7.6	21.3	
		>25mg/dL in 48 hours	796.0	40.0	15.2	5.5	14.0	
		>25mg/dL and AKI in 48 hours	124.7	46.2	13.1	9.6	21.3	
Serum creatinine	μmol/L	Population	2795.3	103.3	56.7	10.9	10.4 [10.4, 10.5]	73.7 [68.2, 78.9]
		AKI in 48 hours	194.4	113.2	40.5	21.0		
		>132.6 μmol/L in 48 hours	479.0	78.0	23.6	11.4		
		>132.6 μmol/L and AKI in 48 hours	129.1	116.5	50.0	21.3		
Serum potassium	mEq/L	Population	2993.4	4.2	0.5	0.3	6.6 [6.6, 6.6]	62.8 [56.0, 68.5]
		AKI in 48 hours	191.1	4.4	0.6	0.4	7.9	
		>5mEq/dL in 48 hours	191.6	5.3	0.2	0.6	6.3	
		>5mEq/dL and AKI in 48 hours	34.7	5.4	0.8	0.7	13.3	
Serum sodium	mEq/L	Population	2995.2	138.2	3.7	1.7	1.2 [1.2, 1.2]	58.9 [41.4, 71.0]
Serum chloride	mEq/L	Population	2939.0	103.6	4.9	2.0	1.9 [1.9, 1.9]	64.4 [16.0, 96.2]
Serum calcium	mEq/L	Population	2576.4	8.8	0.6	0.3	3.0 [2.9, 3.0]	44.8 [39.1, 49.7]
Serum P04	mg/dL	Population	1282.6	3.6	0.9	0.5	14.1 [14.0, 14.2]	62.3 [54.3, 68.7]

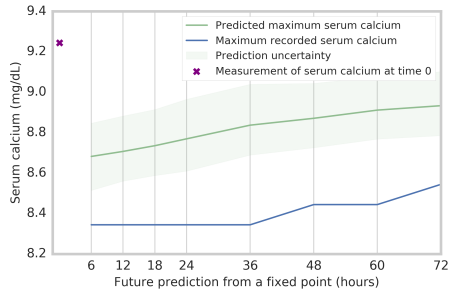
Supplementary Table 3 | Model accuracy in predicting whether a laboratory value will increase in the next 48 hours for a set of seven laboratory test values. When the laboratory test value is substantially increasing (by an amount more than the median increase for that test), the model correctly predicts that the value will increase in 48 hours in 88.5% of cases.

	% predictions correctly predicting an increase in value in 48 hours	
Laboratory test	Cases where the value is increasing	Cases where the value is increasing by an amount more than the median
Serum urea nitrogen	83.7%	90.8%
Serum creatinine	83.6%	86.3%
Serum potassium	85.2%	90.5%
Serum sodium	79.4%	88.5%
Serum chloride	76.9%	86.5%
Serum calcium	84.8%	90.8%
Serum P04	85.2%	91.1%
Weighted average	82.5%	88.5%

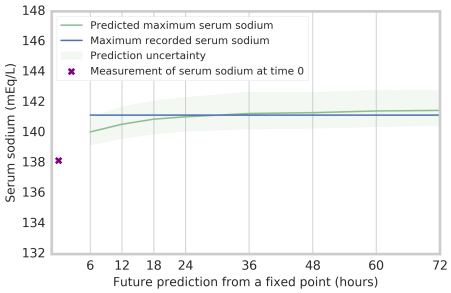


(a) Serum creatinine

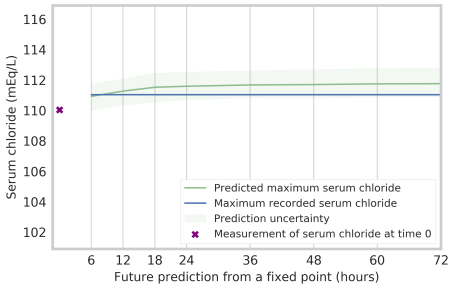
(b) Serum urea nitrogen



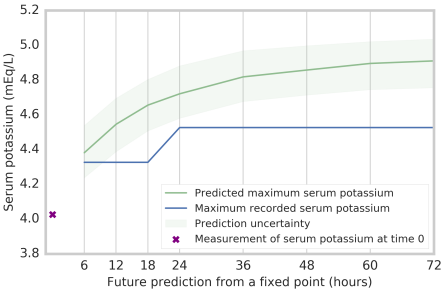
(c) Serum calcium



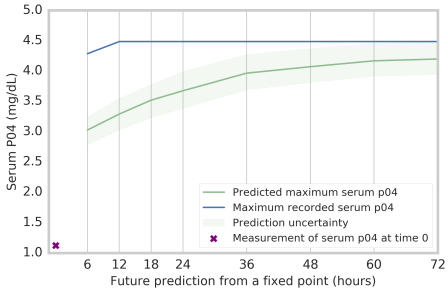
(d) Serum sodium



(e) Serum chloride



(f) Serum potassium



(g) Serum phosphate

Supplementary Figure 2 | Examples of predictions from the auxiliary task. Each figure shows model predictions for the maximum future observed values of a laboratory test value from 6-72 hours in the future from the same fixed point in time, 5 days into a patient admission. The lighter green borders on the prediction curve indicate uncertainty, taken as the range of 100 ensemble predictions once trimmed for the highest and lowest 5 values.

D. Subgroup analysis

The performance of predictive models is not uniform across the entire patient population and understanding how it differs across different clinical subpopulations can help inform choices around future practical deployments.

Supplementary Table 4 outlines differences in PR AUC, ROC AUC, sensitivity and specificity for different subgroups of the VA patient population. PR and ROC AUC do not always increase or decrease at the same time, which is largely due to the differences in the underlying AKI prevalence in different clinical subgroups.

To better understand model performance across different subgroups regardless of the underlying AKI prevalence, we employ error regression. For every observation we computed the expected error given by the logarithmic loss, and fitted a linear regression of the error as an endogenous variable and population subgroups as exogenous variables. A positive computed coefficient points towards a larger model error due to the loss being non-negative. Supplementary Table 5 presents the results of the regression on a subset of predictions with positive primary outcome (AKI of any severity within 48 hours).

In error regression the subgroup performance is modelled jointly, unlike the independent computations of performance presented in Supplementary Table 4. To avoid collinearity in the regression model we removed a set of subgroups corresponding to the most common cases in the data (e.g. age group 50 to 60, unknown ethnicity, male gender, new incoming information in the model, unknown GFR). As the default risk can be taken as constant, the coefficients computed represent a *ceteris paribus* deviation from a default risk for a given subgroup.

The effect of subgroups on the magnitude of errors is jointly significant, as evidenced by F-test (p-value <0.001), as are most of the individual variables corresponding to subgroups. For each such variable this indicates that the magnitude of error is *ceteris paribus* statistically larger/smaller based on the sign than in the default population. For example for admissions with ICU transfers, in the presence of AKI the errors in the model are on average smaller compared

to other admissions. This may suggest either a higher percentage of correct predictions, a higher confidence in making correct predictions, or a lower confidence in making incorrect predictions. This conclusion is supported by the higher PR AUC performance of the models on the ICU transfer patient subpopulation in Supplementary Table 4.

Supplementary Table 4 | Model performance across different clinical subgroups. Performance across multiple clinically important groups when predicting AKI of any severity up to 48 hours ahead of time. Operating points for sensitivity/specificity calculations have been chosen to allow for precision of 33%, which translates to having two false positives for each true positive.

Subgroup name		PR AUC	ROC AUC	Sensitivity (AKI episode)	Sensitivity (step)	Specificity (step)	Positives ratio (step)
Patient demographics	Age group 20-30	11.0%	93.4%	27.5%	18.2%	99.7%	0.39%
	Age group 30-40	20.7%	94.4%	36.7%	22.3%	99.7%	0.58%
	Age group 40-50	18.0%	95.1%	40.8%	24.2%	99.6%	0.62%
	Age group 50-60	26.8%	93.6%	52.6%	33.1%	99.0%	1.35%
	Age group 60-70	31.8%	90.4%	57.6%	36.7%	97.9%	2.75%
	Age group 70-80	31.6%	89.3%	58.2%	36.6%	97.5%	3.15%
	Age group 80-90	28.4%	89.5%	55.7%	32.6%	98.0%	2.76%
	Ethnicity: Black	34.9%	93.9%	60.4%	39.7%	98.5%	1.99%
	Ethnicity: Unknown	28.0%	91.5%	54.1%	33.3%	98.4%	2.09%
	Gender: Female	24.1%	93.1%	44.8%	28.5%	99.2%	1.29%
	Gender: Male	29.9%	92.0%	56.0%	35.1%	98.4%	2.16%
Admissions	Medical admissions	31.1%	88.6%	57.2%	35.7%	97.5%	3.24%
	Surgery admissions	33.2%	88.5%	58.5%	36.5%	97.6%	3.42%
	ICU transfers	36.3%	87.8%	64.3%	40.4%	96.4%	4.68%
	ER visits	30.4%	92.1%	56.7%	34.9%	98.5%	2.00%
	Adm. duration > 7 days	32.4%	93.6%	58.6%	36.0%	98.7%	1.89%
Patients with CKD	All CKD	42.6%	89.3%	70.8%	48.8%	95.1%	5.34%
	CKD stage 1*	18.3%	90.0%	42.8%	22.0%	99.0%	1.52%
	CKD stage 2	24.5%	90.9%	49.3%	29.4%	98.4%	2.19%
	CKD stage 3A	29.3%	86.2%	57.8%	36.4%	95.7%	4.88%
	CKD stage 3B	48.1%	86.1%	73.1%	54.2%	91.4%	8.68%
	CKD stage 4	60.1%	85.8%	83.9%	68.5%	84.1%	13.9%
	CKD stage 5	69.4%	89.2%	85.6%	70.0%	90.4%	13.75%
Other at risk groups	Diabetic patients	32.2%	91.1%	60.3%	39.1%	97.6%	2.88%
	Death within 30 days of adm.	41.8%	90.4%	69.9%	45.3%	96.3%	4.94%
	Death within 7 days of adm.	44.0%	91.1%	71.7%	46.4%	96.3%	5.21%
	Haemoglobin <80g/L	42.3%	88.0%	67.8%	44.2%	96.2%	5.31%
	Haemoglobin <80g/L in the first 2 days	42.0%	87.9%	69.3%	46.4%	95.8%	5.31%
	WCC >12 or <3.5 x10^9/L	33.5%	89.2%	58.9%	36.4%	97.6%	3.44%
	WCC >12 or <3.5 x10^9/L in the first 2 days	32.4%	87.8%	58.0%	36.3%	97.1%	3.82%
	Post IV Contrast administration	33.5%	90.0%	57.0%	34.5%	98.3%	2.68%

*CKD stage 1 is evidence of renal parenchymal damage with a normal glomerular filtration rate (GFR). This is rarely recorded in our dataset; instead the numbers for stage 1 CKD have been estimated from admissions that carried an ICD-9 code for CKD, but where GFR was normal. For this reason these numbers may under-represent the true prevalence in the population.

Supplementary Table 5 | Regression of model errors on population subgroups for N=194,922 positive primary outcomes. The R-squared is 22.9%, and the F-statistic (p-value <0.001) is evidence towards joint significance of the set of 31 covariates.

Variable	Coefficient	Standard deviation	p-value	95% confidence intervals
Default (constant)	3.98	0.02	<0.001	[3.93, 4.03]
Age group 20 to 30	0.64	0.05	<0.001	[0.54, 0.75]
Age group 30 to 40	0.30	0.03	<0.001	[0.24, 0.36]
Age group 40 to 50	0.26	0.02	<0.001	[0.23, 0.30]
Age group 60 to 70	-0.06	0.01	<0.001	[-0.07, -0.04]
Age group 70 to 80	0.01	0.01	0.20	[-0.01, 0.03]
Age group 80 to 90	0.19	0.01	<0.001	[0.17, 0.22]
Ethnicity: Black	-0.14	0.01	<0.001	[-0.15, -0.13]
Gender: Female	0.15	0.02	<0.001	[0.12, 0.19]
Patients with CKD	-0.62	0.01	<0.001	[-0.64, -0.61]
CKD stage 1	0.16	0.01	<0.001	[0.14, 0.18]
CKD stage 2	-0.08	0.01	<0.001	[-0.11, -0.06]
CKD stage 3a	-0.23	0.01	<0.001	[-0.25, -0.21]
CKD stage 3b	-0.56	0.01	<0.001	[-0.59, -0.54]
CKD stage 4	-0.95	0.01	<0.001	[-0.98, -0.93]
CKD stage 5	-1.09	0.03	<0.001	[-1.14, -1.05]
Medical admissions	-0.16	0.01	<0.001	[-0.17, -0.15]
Surgery admissions	-0.19	0.01	<0.001	[-0.20, -0.17]
ICU transfers	-0.31	0.01	<0.001	[-0.33, -0.30]
ER visits	0.09	0.01	<0.001	[0.08, 0.11]
Diabetic patients	-0.11	0.01	<0.001	[-0.12, -0.09]
Death within 30 days of admission	-0.17	0.02	<0.001	[-0.20, -0.14]
Death within 7 days of admission	-0.14	0.02	<0.001	[-0.17, -0.10]
Haemoglobin <80g/L	-0.23	0.01	<0.001	[-0.25, -0.22]
Haemoglobin <80g/L in first 2 days	0.02	0.01	0.11	[-0.00, 0.04]
WCC >12 or <3.5 x10^9/L	-0.01	0.01	0.30	[-0.03, 0.01]
WCC >12 or <3.5 x10^9/L in first 2 days	-0.15	0.01	<0.001	[-0.17, -0.14]
Admission duration > 7 days	0.11	0.01	<0.001	[0.10, 0.13]
Post IV contrast administration	-0.04	0.01	<0.001	[-0.05, -0.03]
Post IV saline administration	-0.23	0.02	<0.001	[-0.27, -0.20]
Old information aggregation only	0.30	0.01	<0.001	[0.29, 0.31]
Admission with at least 1 AKI	-0.93	0.02	<0.001	[-0.97, -0.89]

E. Literature review

E.1. AKI risk models

Supplementary Table 6 | Results from a literature review of papers investigating the risk prediction of AKI

Author/Year	Country	Num. sites	Patient subgroup	Num. patients	Num. admissions	AKI definition	Time of prediction	Independent test set	Best performing model architecture(s)	ROC AUC	Other perf. measures
Drawz 2008 [64]	U.S.	3	Adults admitted to medicine, surgery or obstetrics	540	-	AKIN criteria AKI during admission	Point of admission	Y	Logistic Regression	66%	-
Matheny 2010 [65]	U.S.	1	Adults with admissions of $\geq$2 days duration	21,074	26,107	RIFLE criteria Risk or Injury between days 2 and 30 of admission	Point of admission	N	Logistic Regression	Risk: 75% Injury: 78%	-
Forni 2013 [66]	U.K.	1	Patients admitted to Acute Admissions Unit	1,314	-	KDIGO criteria AKI within 7 days of admission	Point of admission to Acute Admissions Unit	Y	Logistic Regression	72%	-
Cronin 2015 [67]	U.S.	116	Admissions 2-30 days in length	1,620,898	-	KDIGO criteria AKI between days 2 and 9 of admission	48 hours after admission	N	Logistic Regression	AKI Stages 1-3: 76% AKI Stages 2-3: 72%	-
Bedford 2016 [68]	U.K.	3	All admissions	-	775 to 9157[2]	New KDIGO criteria AKI at (i) admission, (ii) 72 hours after admission, (iii) worsening of KDIGO AKI stage for patients with stage 1 or 2 at presentation, 72 hours after admission	(i) Point of admission, (ii) 24 hours after admission, (iii) Point of admission	Y	Logistic Regression	AKI Stages 1-3: 75% AKI Stages 2-3: 75%	-
Kate 2016 [69]	U.S.	15	Patients $\geq$60 years old	17,044	-	New AKIN AKI between 24 hours after hospital discharge[5]	24 hours after admission	N	Logistic Regression, Ensemble	LR: 66% Ensemble: 66%	-
Koyner 2016 [4]	U.S.	5	All adult inpatients	-	202,961	KDIGO AKI within 24 hours[6]	Every 12 hours	Y	Logistic Regression	AKI 1+: 74% AKI 2+: 76% AKI 3: 83%	-
Thottakkara 2016 [70]	U.S.	1	Patients undergoing surgical procedures	50,318	-	KDIGO AKI within 7 days of procedure	Point of procedure	Y	Logistic Regression, Generalised Additive Model	LR: 82% GAM: 83%	LR PPV: 73% GAM PPV: 72%
Cheng 2017 [5]	U.S.	1	Patients aged 18-64 years old	33,703	48,955	KDIGO AKI within 24 hours	Various time points	N	Random Forest, Logistic Regression	RF: 76.5% LR: 76.3%	RF Precision: 69.2% RF Recall: 0.711% LR Precision: 70.4% LR Recall: 71.1%
Davis 2017 [71]	U.S.	All VA hospitals	All admissions 2-30 days in length	-	1,841,951	New KDIGO AKI between 48 hours and 9 days of admission	48 hours after admission	Y	Random Forest	73%	-
Hodgson 2017 [72]	U.K.	1	Adult medical and general surgical admissions	-	12,554	KDIGO AKI within 7 days[7]	Point of hospital admission	N/A[3]	Logistic Regression	**Medical patients:** Baseline: 64% No baseline: 71% **Surgical patients:** Baseline: 66% No baseline: 67%	-
Mohamadlou 2017 [28]	U.S.	2[1]	All patients	-	68,319	NHSE algorithm AKI at various time points before onset	12, 24, 48 and 72 hours before onset	Y	Gradient Boosted Trees	**BIDMC (ITU only):** 12h: 74.9% 24h: 75.8% 48h: 70.7% 72h: 67.4% **SMC (inpatients):** 12h: 80% 24h: 79.5% 48h: 76.1% 72h: 72.8%	**BIDMC (ITU only):** Sens 77%-83% Spec 45%-75% **SMC (inpatients):** Sens 75%-85% Spec 51%-82%
Weisenthal 2017 [73]	U.S..	1	Readmissions	12,491	-	ICD-9 code OR KDIGO AKI during admission	Point of hospital readmission	Y	MLP	92%	PR AUC: 70%
Adhikari 2018 [74]	U.S.	1	Patients undergoing surgery	2,911	-	KDIGO AKI within (i) 3 postoperative days, (ii) 7 postoperative days, and (iii) up to the point of hospital discharge	Before and after index surgery	Y	Random Forest	**Pre-operative models:** 3 day: 83.37% 1 day 84.4% admission: 83.7% **Post-operative models:** 3 day: 84.57% 1 day: 86.0% Admission: 85.4%	**Pre-operative model:** 3 day: Sens: 82.4% Spec: 63.8% PPV: 55.1% NPV: 87%
Bihorac 2018 [18]	U.S.	1	Patients undergoing surgery	51,457	-	RIFLE AKI during admission	Before index surgery	N[6]	Generalised Additive Model	88%	Sens 80% Spec 79% PPV 72% NPV 85% Accuracy 80%
Koyner 2018 [6]	U,S.	1	All patients	-	121,158	KDIGO AKI within 48 hours	First creatinine measurement after admission	Y	Random Forest	AKI Stages 1-3: 73% AKI Stages 2-3: 87% AKI Stage 3: 93%	NPV and PPV presented for a variety of predicted probability cut-offs
Park 2018 [75]	Korea	1	Cancer patients	21,022	-	Adjusted baseline KDIGO AKI within 14 days	Inpatient creatinine measurement	Y	Random Forest	-	Precision: 78.9% Recall: 75.1% F-measure: 75.8%
Weisenthal 2018 [76]	U.S.	1	Re-admissions	34,505	-	ICD-9 code OR KDIGO during admission	Point of hospital re-entry	Y	Gradient Boosted Trees	86.7%	PR AUC: 32.6%
Li 2018 [77]	U.S.	1	ICU patients	$\sim$40,000	-	KDIGO	24h after admission	Y	Convolutional Neural Network	77.9%	Precision: 40.7% Recall: 65.4%
Pan 2019 [29]	U.S.	1	ICU patients	40,000	58,000	RIFLE AKI during admission	Inpatient Various time points	Y	Recurrent neural network	-	ROC AUC: 88.9% and 83.7%

[1] ITU only (BIDMC) and Inpatients (SMC); [2] Model dependent; [3] External validation of Forni 2013; [4] TRIPOD 1b; [5] Excluded those with diagnosis of AKI within 24 hours of admission and those with CKD stage 3-5

[6] Discrete time survival model. Excluded patients with initial SCr >3mg/dl or who developed AKI prior to ward admission; [7] Excluded patients admitted to ITU from ED

E.2. Literature: Machine Learning Models for EHR

There has been significant recent progress in applications of machine learning to modelling clinical data based on electronic health records [78]. We provide a systematic overview of these achievements in Supplementary Table 7. Machine learning models have shown promise when used for predicting mortality [3, 9, 79], sepsis [10, 70, 80], post-operative complications [18, 81], readmission risk [11], for providing treatment recommendations [82], modelling treatment response [15, 32], detecting early signs of heart failure [83–85] and in planning for palliative care [8]. Most of the deep learning approaches involve improvements in representation learning [12] or apply recurrent neural networks (RNN) [9, 10, 13, 83, 86, 87] or convolutional models [11, 14, 35, 88].

Despite these recent advances, building robust clinically applicable risk models from routinely collected EHR data remains a challenge [89]. Clinically applicable models need to be able to reliably deliver personalised insights on preventable conditions, early enough to enable clinical intervention and providing enough information to inform decision making. Models need to be evaluated on large representative datasets and be capable of integrating all of the available relevant medical information. The evaluation needs to be performed with the application in mind, and good levels of sensitivity need to be achieved under clinically applicable levels of precision. These challenges provide a barrier to implementation.

Supplementary Table 7 | Results from a literature review of papers proposing machine learning models for modelling electronic health records

Author/Year	Num. patients	Num. admissions	Num. features	Clinical tasks	Model architecture
Lim 2018 [9]	10,980	-	87	Mortality, cystic fibrosis, comorbidities	LSTM + additional layers
Rajkomar 2018 [3]	114,003	216,221	all available data	Mortality, readmission, long length of stay, discharge diagnosis	LSTM, TANN, boosted decision stumps
Futoma 2017 [10]	-	49,312	77	Sepsis	GP + LSTM
Nguyen 2016 [11]	∼300,000	590,546	diagnoses, procedures	Readmission	CNN
Wang 2018 [82]	∼43,000	22,865	-	Treatment optimisation	SRL-RNN
Avati 2017 [8]	221,284	-	13,654	(3-12 month) Mortality	MLP
Miotto 2016 [12]	∼700,000	-	41,072	Disease prediction	stacked denoising AEs
Lipton 2017 [13]		10,401	13	Diagnosis classification	LSTM
Choi 2016 [86]	263,706	-	1,778	Predicting properties of subsequent visits	GRU
Choi 2016 [83]	32,787	-	diagnoses, procedures, medication	Heart failure detection	GRU
Che 2016 [87]	-	58,000	99	Mortality, diagnosis category	GRU-D
Razavian 2016 [88]	∼298,000	-	44	CKD progression	CNN, LSTM
Cheng 2016 [14]	319,650	-	diagnoses	Congestive heart failure, chronic obstructive pulmonary syndrome	CNN
Komorowski 2018 [7]	96,156	-	48	Sepsis treatment	MDP
Henao 2016 [79]	240,000	4,400,000	24,567	Mortality and morbidity	Deep Poisson factor models
Soleimani 2017 [15]	67	-	5	Dialysis treatment response	Gaussian processes
Schulam 2017 [32]	428	-	4	Dialysis treatment response	Gaussian processes
Alaa 2016 [16]	6,313	-	12	Risk of adverse events	Hierarchical latent class model and Gaussian processes
Thottakkara 2016 [70]	50,318	-	285	Post-operative AKI and sepsis	Naive Bayes and SVM
Bihorac 2018 [18]	51,457	-	-	Post-operative complications	Generalised additive model
Perotte 2015 [17]	2,908	-	106	CKD progression	Kalman filter and Cox proportional hazards
Hu 2015 [81]	6,258	-	demographics, diagnoses, orders, labs, vitals, medications	Surgical site infections	Logistic regression
Sideris 2015 [84]	3,041	-	demographics, diagnoses, labs	Heart failure	SVM + clustering
Goldstein 2014 [85]	1,718	-	72	Sudden cardiac death	Random forests
Mani 2014 [80]	299	1826	811	Neonatal sepsis	Random forests SVM CART Logistic regression
Henry 2015 [2]	16,234	-	54	Sepsis	Cox proportional hazards model

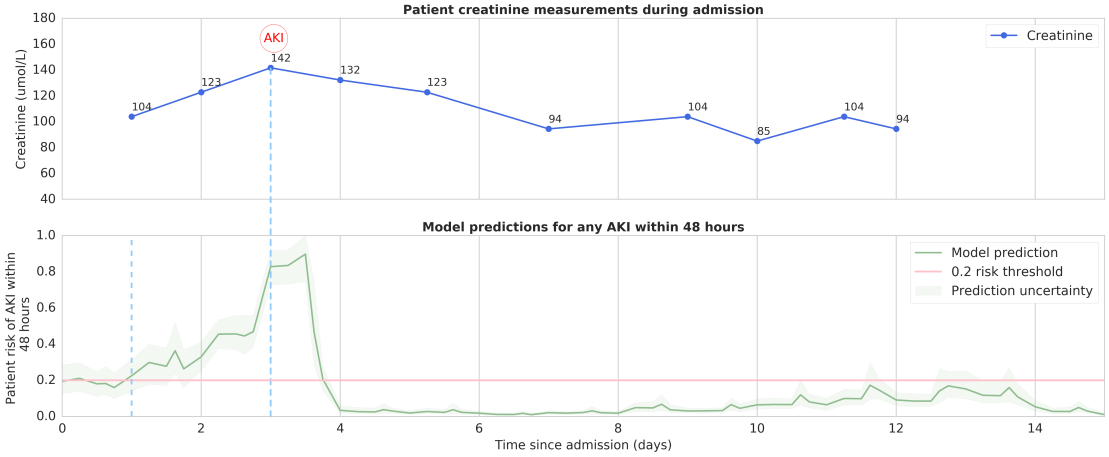
F. Success and failure cases

To demonstrate examples of how the model perceives the risk of AKI during an admission we provide a visual representation in Figure 1 in the main text that this supplementary material accompanies.

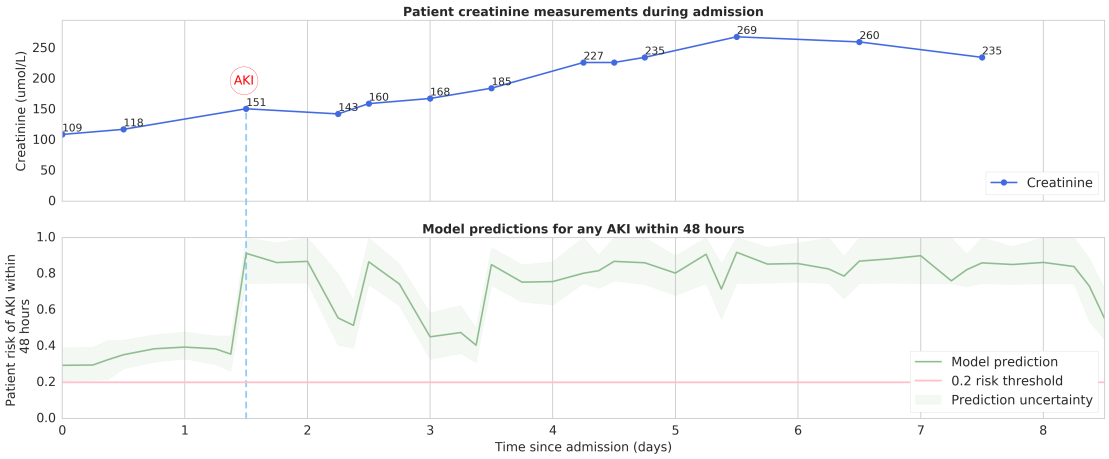
To avoid demonstrating the performance of the model by 'cherry picking' a single example, we present an additional set of five systematically selected success and failure cases of the predictive model. In each of these examples, the first plot shows the creatinine measurements throughout the admission from the EHR, and the second plot shows the model's continuous risk predictions from an ensemble of 100 predictive models. In each case the risk curve represents the mean prediction across the ensemble and the lighter green borders on the risk curve indicate uncertainty, taken as the range of 100 ensemble predictions once trimmed for the highest and lowest 5 values.

These cases were selected systematically as the 'best' success cases, maximising first for the number of correct positive predictions and then for correct negative predictions while allowing at most one incorrect prediction, and the 'worst' failure cases, maximising for the number of false positive or false negative predictions during an admission. They were selected after filtering out examples where renal replacement therapy had occurred prior to an AKI, or where severe CKD had been recognised prior to an AKI.

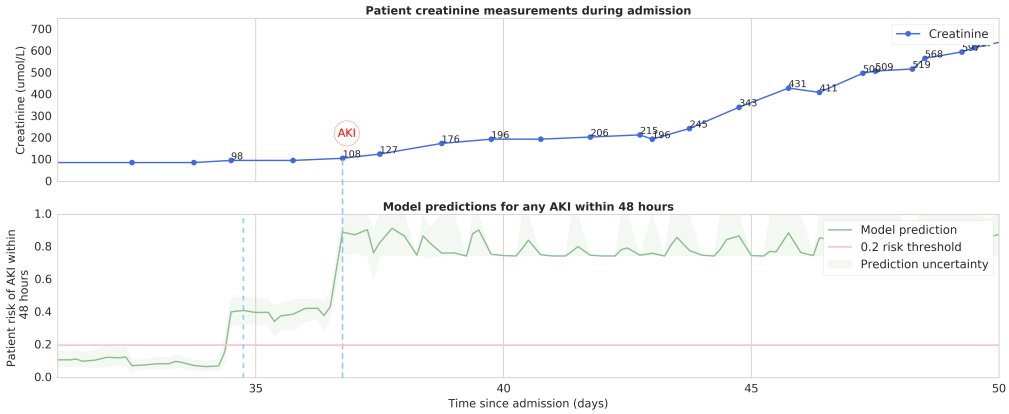
F.1. Success case examples



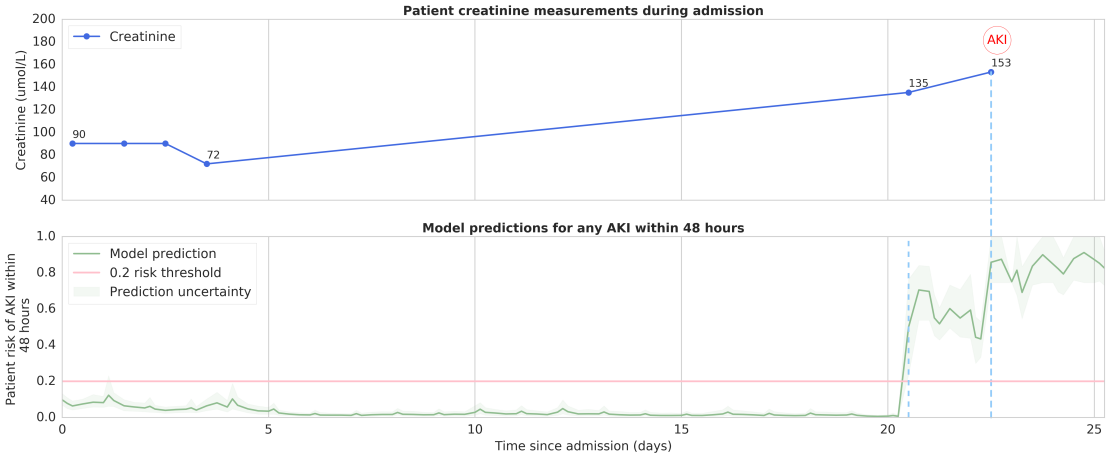
Supplementary Figure 3 | Visual representation of a 15 day surgical admission for a 77 year old male patient with a history of congestive heart failure. The patient developed AKI 3 days after admission, with accompanying evidence of sepsis. The model correctly predicts the patient is at risk 48 hours before the AKI is detected according to KDIGO criteria.



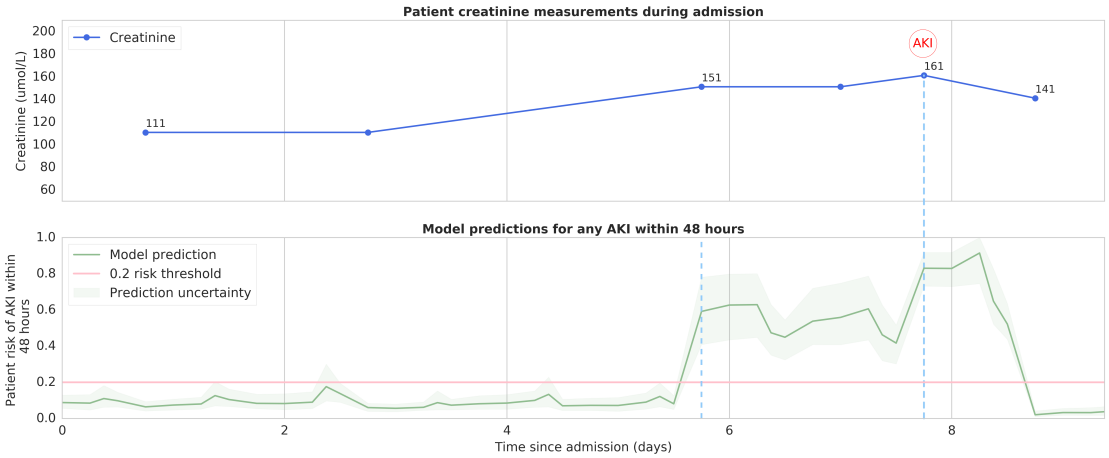
Supplementary Figure 4 | Visual representation of a 9 day intensive care admission for a 57 year old male with a history of diabetes. The first onset of AKI occurs during the second day of admission; from the beginning of the admission the model predicts the risk at above the 0.2 threshold. Ultimately the patient went on to develop chronic kidney disease after discharge.



Supplementary Figure 5 | A 19 day section of an 8 week admission of a 59 year old male with past history of diabetes. Despite normal renal function, the model correctly predicts an impending AKI, 48 hours before the event occurs on the 36th day of admission. The AKI progressed to require an intensive care admission and haemofiltration; the patient passed away at the end of admission.

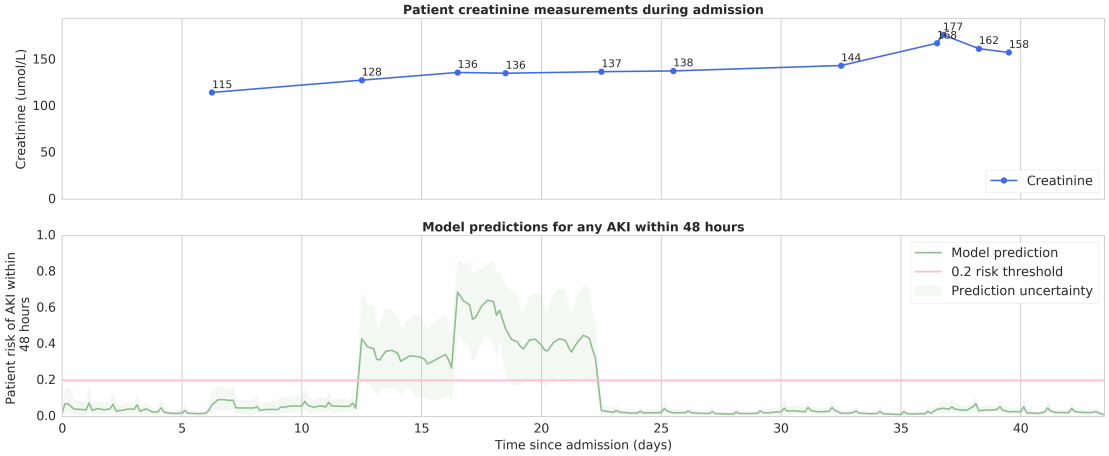


Supplementary Figure 6 | Visual representation of an admission under the medical team of a 64 year old male with a history of CKD and congestive heart failure. After a long period without blood measurements, the patient developed an AKI on the 22nd day of admission, which was correctly anticipated by the model.

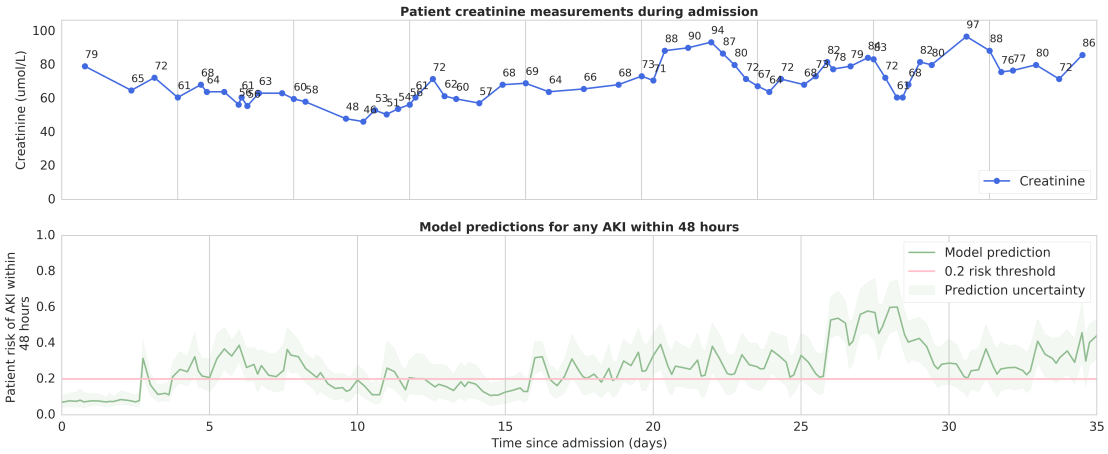


Supplementary Figure 7 | A visual representation of a 10 day medical admission of a 60 year old male with a history of congestive heart failure. The model correctly predicts the gradual increase of creatinine being labelled as AKI by KDIGO criteria.

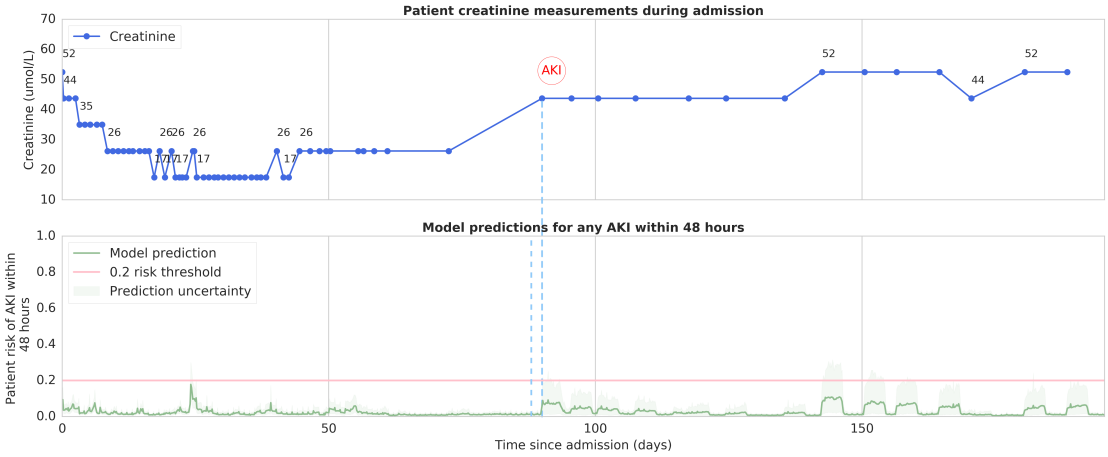
F.2. Failure case examples



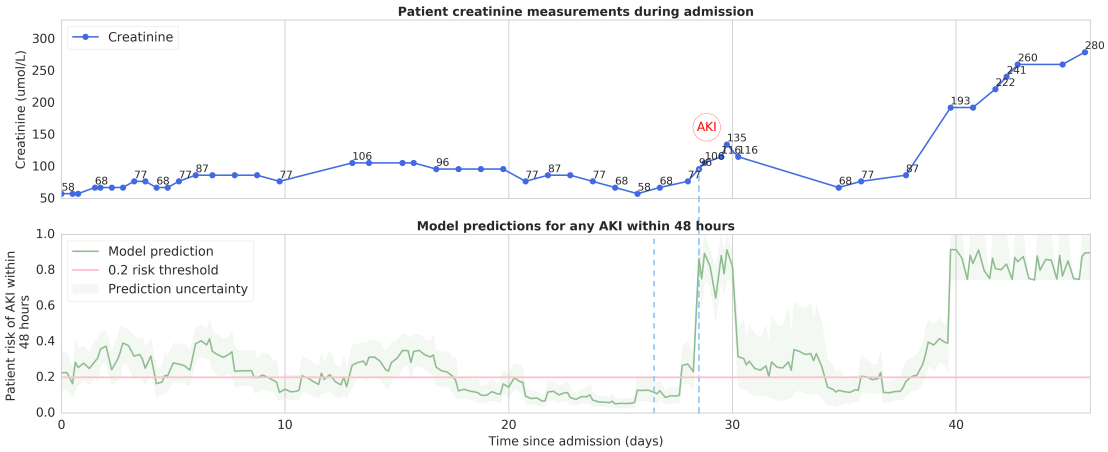
Supplementary Figure 8 | A 59 year old male with a history of CKD, admitted under the medical team with evidence of sepsis and transferred to the intensive care unit 2 days after admission. Despite infrequent creatinine measurements in the patient records, e-GFR is consistently measured, suggesting information is missing in the records. The model incorrectly suggests a raised risk of AKI during the admission which was not followed by an AKI event, though later on in the admission the creatinine rises well above the patients pre-admission baseline levels. Due to the longer period over which the creatinine has increased, the KDIGO calculated baseline has adjusted and this event is no longer labelled as an AKI event in the dataset.



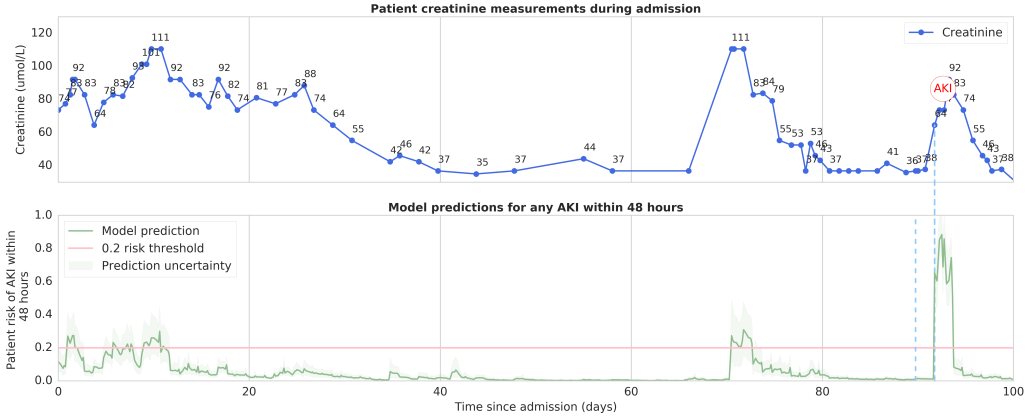
Supplementary Figure 9 | A 57 year old male with multiple previous AKI episodes in previous admissions, admitted here with evidence of infection. Despite a long 35 day admission with frequently raised inflammatory markers the patients renal function remained stable; the model provides raised risk scores throughout this admission.



Supplementary Figure 10 | A lengthy 27 week admission of a 45 year old male with a history of diabetes, admitted directly admitted into the intensive care unit. The patient has a consistently low creatinine, possibly due to low muscle mass, which results in a rise from 26 to 44 μmol/L over several weeks being categorised by KDIGO criteria as an AKI. While cases such as this are reported in our results as false negative predictions, the clinical relevance of such a failure is negligible.



Supplementary Figure 11 | A 64 year old male with a history of chronic obstructive pulmonary disease (COPD) and diabetes, admitted directly to intensive care with evidence of an infective exacerbation of COPD. The patient was transferred to intensive care two further times during the six week admission. The model incorrectly provides a raised risk of AKI during the early stages of the admission; however the first AKI event occurs much later on day 28 which is then correctly predicted by the model, 18 hours ahead of time. Though this resolves a more severe AKI occurs later in the admission. The patient ultimately deteriorates and passes away during this inpatient stay.



Supplementary Figure 12 | The first 100 days of another lengthy admission, this time lasting 7 months. A 73 year old male with a history of diabetes is admitted directly to the intensive care unit. The model raises the risk of AKI early on in the admission, and though this is accompanied by an increase from 60 to 111 μmol/L of creatinine, the duration over which it increases does not meet KDIGO criteria. Much later on in the admission, similar rises occur where the model does not provide a proactive increase in risk. The second of these meets KDIGO criteria.

G. Hyperparameter sweeps

Finding the best AKI risk model architecture was an iterative process that involved trying different design choices and model parameters and evaluating the model performance on the validation set. This resulted in the final set of parameters reported in Methods. The full range of hyperparameter options considered in our experiments during the model development process is displayed in Supplementary Table 8.

Supplementary Table 8 | Hyperparameter combinations evaluated in the experiments

Hyperparameter	Values considered
RNN cell type	LSTM, GRU, UGRNN, SRU, Intersection RNN, MANN, NTM, DNC, RMC
RNN cell size	100, 150, 200, 250, 300, 400, 500
RNN num. layers	1, 2, 3
Embedding num. layers	1, 2, 3
Embedding dim. per feature type	200, 250, 300, 400, 500
Embedding combination	concatenate, sum
Embedding architecture type	MLP, AE, VAE
Embedding reconstruction loss weight	1e-2, 1e-3, 1e-4
Embedding reconstruction sampling ratio	1, 2, 5, 10
Optimise directly for PR AUC	on, off
Highway connections	on, off
Residual embedding connections	on, off
Input dropout	0, 0.1, 0.2, 0.3
Output dropout	0, 0.1, 0.2, 0.3
Embedding dropout	0, 0.1, 0.2, 0.3
Variational dropout	0, 0.1, 0.2, 0.3
Input regularisation type	None, L1, L2
Input regularisation term weight	1e-3, 1e-4, 1e-5
BPTT Window	32, 64, 128, 256, 512
Embedding activation functions	Tanh, ReLU [90], Leaky ReLu [91], Swish [92], ELU [93], SELU [94], ELiSH [95], Hard ELiSH [95], Sigmoid, Hard Sigmoid
Auxiliary task loss weight	0., 0.1, 0.5, 1, 5, 10
Learning rate	1e-2, 1e-3, 1e-4, 1e-5
Learning rate decay scheduling	on, off
Learning rate decay num. steps	6000, 8000, 12000, 15000, 20000
Learning rate decay base	0.7, 0.8, 0.85, 0.9, 0.95
Batch size	32, 64, 128, 256, 512
NTM/DNC memory capacity	64, 128, 256
NTM/DNC memory word size	16, 32, 64
NTM/DNC memory num. reads	6, 10
NTM/DNC memory num. writes	1, 2, 3

H. Model comparison

We have conducted a broad comparison of available models on the AKI prediction task. We considered three broad classes of models and found that:

- Recurrent neural networks (SRU, NTM, LSTM, MANN, DNC, UGRNN, GRU, Intersection RNN, RMC) achieve the highest performance for both PR AUC and ROC AUC, with minimal difference between each other. They also require the fewest training features: they are able to achieve the same performance only with sequential information and the last 48 hours of patient history and can aggregate the patient information while traversing the sequence.

- Feed-forward models (deep MLP, shallow MLP, Logistic Regression, Random Forest, Gradient Boosted Trees) do not have the capacity to aggregate the information about a patient over time, which necessitates manual collection and engineering of patient historical features. In these models we have experimented with using either 6 months of 5 years of historical information and we are reporting the better performing of the two for each.

- Gradient Boosted Trees (GBTs) benefited from heavy overweighting of observations with positive-labels while equivalent oversampling for random forest and neural-network-based models did not bring a similar improvement.

- Since tree-based methods are batch methods that cannot fit all data in memory – and online variants typically underperform standard ones – they were trained on one-third of the patient data. To establish whether training these baselines on a third of the training data had an adverse impact on performance, we conducted experiments to assess how the model performance changes upon further reduction. A further reduction in the number of patients in the training data of 40% resulted in only minor changes in ROC AUC and PR AUC which degraded by 0.2% and 0.8% respectively. This suggests that potential minor improvements in the tree baseline performance could have been obtained if it had been

possible to provide the entirety of the data, but that these would have still fallen short of the RNN performance by a large margin.

Supplementary Table 9 | Comparison of different predictive models and RNN cells. *SRU significantly outperforms the Logistic Regression, Gradient Boosted Trees and Random Forest baselines in terms of PR AUC for the main task of predicting any AKI up to 48 hours ahead of time; using two-sided Mann–Whitney U test on 200 samples per model (see Evaluation) SRU is significantly better with a p-value of <0.001.

AKI task	Model	PR AUC (%) [95% CI]	ROC AUC (%) [95% CI]
Any AKI	SRU	29.7 [28.5, 30.8]	92.1 [91.9, 92.3]
up to 48 hours early	Intersection RNN	29.6 [28.5, 30.7]	91.9 [91.7, 92.1]
	NTM	29.0 [27.6, 30.0]	91.9 [91.5, 91.9]
	MANN	28.9 [27.8, 30.0]	92.0 [91.8, 92.2]
	LSTM	28.8 [27.7, 30.0]	92.1 [91.8, 92.2]
	UGRNN	28.3 [27.2, 29.5]	91.9 [91.7, 92.1]
	GRU	27.8 [26.7, 28.8]	92.0 [91.8, 92.2]
	RMC	26.2 [25.0, 27.3]	91.3 [91.1, 91.5]
	DNC	26.5 [25.4, 27.4]	91.9 [91.7, 92.1]
	Deep MLP	25.1 [23.9, 26.1]	90.3 [90.0, 90.6]
	CNN	23.8 [22.8, 24.8]	90.1 [89.9, 90.4]
	Shallow MLP	22.3 [21.1, 23.2]	89.9 [89.6, 90.1]
	Gradient Boosted Trees*	22.0 [21.0, 22.9]	88.9 [88.6, 89.2]
	Random Forest*	19.8 [18.8, 20.9]	87.1 [86.7, 87.4]
	Logistic Regression*	17.3 [16.2, 18.2]	86.3 [86.0, 86.7]
AKI	Intersection RNN	37.8 [35.7, 40.0]	95.7 [95.5, 96.0]
stages 2 and 3	UGRNN	37.3 [35.1, 39.2]	95.6 [95.3, 95.9]
up to 48 hours early	LSTM	37.1 [35.4, 39.1]	95.5 [95.2, 95.8]
	NTM	36.9 [35.1, 39.0]	95.5 [95.2, 95.7]
	GRU	36.2 [34.2, 38.1]	95.5 [95.2, 95.8]
	MANN	36.2 [34.6, 38.1]	95.4 [95.1, 95.7]
	DNC	35.7 [33.6, 37.5]	95.5 [95.2, 95.8]
	Deep MLP	32.2 [30.2, 33.9]	94.9 [94.5, 95.2]
	SRU	29.0 [27.1, 30.6]	94.7 [94.4, 95.0]
	CNN	27.2 [25.3, 28.9]	94.3 [93.9, 94.6]
	Shallow MLP	25.3 [23.9, 26.8]	93.7 [93.4, 94.1]
	Gradient Boosted Trees	25.1 [23.3, 26.8]	92.5 [92.2, 92.9]
	Random Forest	25.1 [22.9, 26.6]	91.1 [90.6, 91.5]
	RMC	21.9 [20.5, 23.2]	91.1 [90.6, 91.6]
	Logistic Regression	16.7 [15.2, 18.1]	87.0 [86.3, 87.6]
AKI	NTM	48.7 [46.4, 51.1]	98.0 [97.8, 98.2]
stage 3	MANN	47.9 [45.8, 50.0]	98.0 [97.7, 98.1]
up to 48 hours early	Intersection RNN	47.8 [45.3, 50.2]	98.0 [97.8, 98.2]
	GRU	47.5 [45.6, 49.9]	98.0 [97.8, 98.2]
	UGRNN	47.1 [45.1, 49.1]	98.1 [97.9, 98.2]
	LSTM	46.8 [44.7, 49.3]	98.0 [97.8, 98.2]
	SRU	46.6 [44.4, 48.9]	98.0 [97.8, 98.2]
	DNC	45.0 [42.0, 47.5]	97.8 [97.6, 98.0]
	Deep MLP	40.9 [38.8, 42.9]	97.5 [97.3, 97.8]
	CNN	38.8 [36.8, 41.0]	97.3 [97.1, 97.5]
	Random Forest	34.6 [31.9, 37.2]	95.5 [95.2, 95.9]
	Gradient Boosted Trees	32.9 [30.9, 35.0]	96.2 [95.9, 96.5]
	Shallow MLP	32.7 [30.8, 34.6]	96.7 [96.4, 96.9]
	RMC	24.7 [22.2, 26.4]	93.8 [93.3, 94.3]
	Logistic Regression	24.5 [23.1, 25.9]	93.0 [92.5, 93.6]

I. Ablation study

We analyse the contribution of the aspects of our model's design to its overall performance, conducting an ablation study that removes specific components of the model, training it fully, and then comparing the simplified model's PR AUC on the validation set. We show the result of this analysis in Supplementary Table 10. We investigate the effect of making the input embeddings shallow, i.e. only using one neural network layer instead of several. We also inspect the effect of removing embedding regularisation. In all cases we see a non-trivial reduction in performance when each of these components are removed. The removal of the auxiliary prediction loss and the removal of regularisation resulted in some of the largest drops in model performance.

We also compare models trained on only the sequential information to models augmented with historical features over short-term (last 48 hours) and long-term (last 6 months) time frames. The results are presented in Supplementary Table 11. The RNN model is able to aggregate information across time and there is a smaller difference in performance than for logistic regression which benefits heavily from hand-crafted historical features.

Supplementary Table 10 | Model performance with ablations. Performance is expressed in PR AUC. We compare the performance for a recurrent model (SRU) and feed-forward model (MLP) on predicting any AKI within 48 hours. 95% confidence intervals are calculated from an un-paired z-test, with 50 models trained from random initialisation per configuration.

PR AUC	**SRU**	**MLP**
Full model	29.7 ± 1.2	25.1 ± 1.1
Shallow model	23.1 ± 0.7	22.9 ± 0.1
Without regularisation	22.5 ± 1.3	23.3 ± 0.1
Without auxiliary regression	26.6 ± 1.4	24.3 ± 0.1
Without numerical features	20.6 ± 0.6	16.7 ± 0.5
Without presence features	22.4 ± 0.9	18.6 ± 0.2

Supplementary Table 11 | Model PR AUC performance for models using sequential and short-term information and optionally being augmented with long-term history aggregation.

PR AUC [95% CI]	**Intersection RNN**	**Logistic Regression**
Sequential information only	28.5 [27.3, 29.4]	14.7 [13.9, 15.4]
Sequential + historical aggregations	28.7 [27.5, 29.7]	17.3 [16.3, 18.1]

J. Influence of data recency on model performance

Making correct predictions of the risk of future AKI is not always possible based on the routinely available data and there will be cases where the models do not have access to the information that is needed to make reliable predictions.

For the models to be able to correctly identify developing AKI, the relevant physiological markers need to be available at the critical point when the predictions are being made. If the signal is absent from the EHR, the model can potentially miss cases of AKI that could have otherwise been detected had the relevant blood tests been taken.

To quantify this effect in our experiments, we compare the average volume and recency of data in cases when the model was correctly predicting future AKI to cases in which it missed predicting future AKI episodes (Supplementary Table 12). We compare the availability of the data in 12 and 24 hours prior to the true positive and false negative predictions. The results strongly suggest that the model errors occur more often when there is less data available to inform the model. This implies that one way of further improving the performance of the current predictive models would be to improve the frequency of measurements for the most relevant biochemical tests in those patients that are known to be at a generally higher risk of developing AKI in the future.

Supplementary Table 12 | Influence of data recency on model performance. Comparison of performance for the mean number of EHR entries and the mean number of creatinine measurements in the clinical data available to the model at prediction time for true positive (N=7,140) versus false negative (N=12,391) predictions made prior to the first AKI in an admission. The mean number of entries in the 24 hours prior to prediction is lower for false negative predictions than for true positive predictions using a 2-sided T-test. The mean number of creatinine measurements in the prior 24 hours is also lower for false negative predictions than for true positive predictions using a 2-sided T-test. The results suggest that the model errors occur more often when there is less data available to inform the model.

		True positives		False negatives		
Entry type	Time before prediction	Mean number of entries	95% Confidence interval	Mean number of entries	95% Confidence interval	p-value
All entries	$\leq$ 12 hours	135.0	[134.5, 136.2]	105.5	[105.3, 106.0]	< 0.01
All entries	$\leq$ 24 hours	206.3	[205.2, 207.5]	168.8	[168.3, 169.3]	< 0.01
Serum creatinine	$\leq$ 12 hours	0.83	[0.82, 0.84]	0.64	[0.64, 0.65]	< 0.01
Serum creatinine	$\leq$ 24 hours	1.25	[1.24, 1.26]	1.00	[1.00, 1.01]	< 0.01

K. Clinically relevant feature set for the baselines

We compared our performance to baseline models trained on features that have been chosen by clinicians as being relevant for modelling kidney function. The initial set of clinically relevant features was chosen on the consensus opinion of six clinicians: three senior attending physicians with over twenty years expertise, one from nephrology and two from intensive care; and three clinical residents with expertise in nephrology, internal medicine and surgery. This set of features was further extended by 36 additional features that were discovered as relevant by our deep learning model, in order to further improve the predictive power of the baseline model.

The following features form the final clinically relevant feature set:

- Demographic information (age, gender, ethnicity);
- Admission information (admission from the Emergency Room, medical or surgical admission, transfer to ICU);
- Vital sign measurements (pulse, systolic and diastolic blood pressure, respiratory rate, oxygen saturation);
- Logical Observation Identifiers Names and Codes (LOINC) for specific laboratory tests (serum creatinine, urea nitrogen, estimated GFR, serum potassium, serum sodium, serum phosphate, serum chloride, serum calcium, haemoglobin, haematocrit, haemoglobin A1C, white cell count, Westergren (ESR), C-reactive protein, total serum protein, serum albumin, serum alkaline phosphatase, serum glutamic pyruvic transaminase, serum glutamic-oxaloacetic transaminase, serum direct bilirubin, serum total bilirubin, serum glucose, serum CO2, serum anion gap, serum vancomycin level, arterial blood gas pH, creatine kinase, 24hr urinary protein);
- ICD-9 subcodes for acute and chronic conditions directly associated with an increased risk of AKI (sepsis, dehydration/hypovolaemia, haemorrhage, liver disease, renal tract obstruction, prior AKI, hypertension, chronic or end-stage renal disease, renal cancer, renal transplant, myocardial infarction, diabetes, vascular disease, gout, congestive cardiac

failure, cardiac arrest, Chronic Obstructive Pulmonary Disease);

- Selected medications (intravenous contrast, intravenous saline, non-steroidal anti-inflammatories, diuretics, angiotensin-converting enzyme (ACE) inhibitors, angiotensin receptor blockers (ARB), aminoglycoside antibiotics, beta lactam antibiotics, glycopeptide antibiotics, quinolone antibiotics, cephalosporin antibiotics, certain chemotherapeutic agents, calcineurin inhibitors, proton pump inhibitors, H2 receptor antagonists, selected antivirals, cyanocobalamin, calcitriol, bisphosphonates, phosphate binders, calcium, methotrexate, sulfonamides, paracetamol, acetylcysteine);
- CPT codes associated with haemodialysis/haemofiltration.

In contrast, the entire feature set available in the EHR totals 366 856 distinct features corresponding to different types of entries. One of the advantages of deep learning models in general is that they are capable of automatically determining which are the relevant features for any predictive task.

References

[1] R. Thomson, D. Luettel, F. Healey, and S. Scobie, "Safer care for the acutely ill patient: Learning from serious incidents," *National Patient Safety Agency*, 2007.

[2] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[3] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. Howell, C. Cui, G. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, 2018.

[4] J. L. Koyner, R. Adhikari, D. P. Edelson, and M. M. Churpek, "Development of a multicenter ward based AKI prediction model," *Clinical Journal of the American Society of Nephrology*, pp. 1935–1943, 2016.

[5] P. Cheng, L. R. Waitman, Y. Hu, and M. Liu, "Predicting inpatient acute kidney injury over different time horizons: How early and accurate?," in *AMIA Annual Symposium Proceedings*, vol. 2017, p. 565, American Medical Informatics Association, 2017.

[6] J. L. Koyner, K. A. Carey, D. P. Edelson, and M. M. Churpek, "The development of a machine learning inpatient acute kidney injury prediction model," *Critical Care Medicine*, vol. 46, no. 7, pp. 1070–1077, 2018.

[7] M. Komorowski, L. A. Celi, O. Badawi, A. Gordon, and A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, pp. 1716–1720, 2018.

[8] A. Avati, K. Jung, S. Harman, L. Downing, A. Y. Ng, and N. H. Shah, "Improving palliative care with deep learning," *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 311–316, 2017.

[9] B. Lim and M. van der Schaar, "Disease-Atlas: Navigating disease trajectories with deep learning," *Proceedings of Machine Learning Research*, vol. 85, 2018.

[10] J. Futoma, S. Hariharan, and K. A. Heller, "Learning to detect sepsis with a multitask gaussian process RNN classifier," in *Proceedings of the International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), pp. 1174–1182, 2017.

[11] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A convolutional net for medical records," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 22–30, 2017.

[12] R. Miotto, L. Li, B. Kidd, and J. T. Dudley, "Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, no. 26094, 2016.

[13] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," *International Conference on Learning Representations*, 2016.

[14] P. Z. J. H. Yu Cheng, Fei Wang, "Risk prediction with electronic health records a deep learning approach," in *Proceedings of the SIAM International Conference on Data Mining*, pp. 432–440, 2016.

[15] H. Soleimani, A. Subbaswamy, and S. Saria, "Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions," *arXiv Preprint arXiv:1704.02038*, 2017.

[16] A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized risk scoring

for critical care patients using mixtures of gaussian process experts," *arXiv Preprint arXiv:1605.00959*, 2016.

[17] A. Perotte, N. Elhadad, J. S. Hirsch, R. Ranganath, and D. Blei, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," *Journal of the American Medical Informatics Association*, vol. 22, no. 4, pp. 872–880, 2015.

[18] A. Bihorac, T. Ozrazgat-Baslanti, A. Ebadi, A. Motaei, M. Madkour, P. M. Pardalos, G. Lipori, W. R. Hogan, P. A. Efron, F. Moore, *et al.*, "MySurgeryRisk: Development and validation of a machine-learning risk algorithm for major complications and death after surgery," *Annals of Surgery*, 2018.

[19] A. Khwaja, "KDIGO clinical practice guidelines for acute kidney injury," *Nephron Clinical Practice*, vol. 120, no. 4, pp. c179–c184, 2012.

[20] C. Stenhouse, S. Coates, M. Tivey, P. Allsop, and T. Parker, "Prospective evaluation of a modified early warning score to aid earlier detection of patients developing critical illness on a general surgical ward," *The British Journal of Anaesthesia*, vol. 84, no. 5, p. 663P, 2000.

[21] J. L. Alge and J. M. Arthur, "Biomarkers of AKI: A review of mechanistic relevance and potential therapeutic implications," *Clinical Journal of the American Society of Nephrology*, vol. 10, no. 1, pp. 147–155, 2015.

[22] H. E. Wang, P. Muntner, G. M. Chertow, and D. G. Warnock, "Acute kidney injury and mortality in hospitalized patients," *American Journal of Nephrology*, vol. 35, pp. 349–355, 2012.

[23] A. MacLeod, "NCEPOD report on acute kidney injury—must do better," *The Lancet*, vol. 374, no. 9699, pp. 1405–1406, 2009.

[24] M. E. Thomas, C. Blaine, A. Dawnay, M. A. Devonald, S. Ftouh, C. Laing, S. Latchem, A. Lewington, D. V. Milford, and M. Ostermann, "The definition of acute kidney injury and its use in practice," *Kidney International*, vol. 87, no. 1, pp. 62 – 73, 2015.

[25] W. A. Cheungpasitporn and K. Kashani, "Electronic data systems and acute kidney injury," *Contributions to Nephrology*, vol. 187, pp. 73–83, 2016.

[26] F. P. Wilson, M. G. S. Shashaty, J. M. Testani, I. Aqeel, Y. Borovskiy, S. S. Ellenberg, H. I. Feldman, H. E. Fernandez, Y. Gitelman, J. Lin, D. Negoianu, C. R. Parikh, P. P. Reese, R. Urbani, and B. D. Fuchs, "Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial," *The Lancet*, vol. 385, pp. 1966–1974, 2015.

[27] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, 2016.

[28] H. Mohamadlou, A. Lynn-Palevsky, C. Barton, U. Chettipally, L. Shieh, J. Calvert, N. R. Saber, and R. Das, "Prediction of acute kidney injury with a machine learning algorithm using electronic health record data," *Canadian Journal of Kidney Health And Disease*, vol. 5, 2018.

[29] Z. Pan, H. Du, K. Yuan Ngiam, F. Wang, P. Shum, and M. Feng, "A self-correcting deep learning approach to predict acute conditions in critical care," *arXiv Preprint arXiv:1901.04364*, 2019.

[30] S. Park, S. H. Baek, S. Ahn, K.-H. Lee, H. Hwang, J. Ryu, S. Y. Ahn, H. J. Chin, K. Y. Na, D.-W. Chae, and S. Kim, "Impact of electronic acute kidney injury (AKI) alerts with automated nephrologist consultation on detection and severity of AKI: A quality improvement study," *American Journal of Kidney Diseases*, vol. 71, no. 1, pp. 9–19, 2018.

[31] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?," *arXiv Preprint arXiv:1805.12002*, 2018.

[32] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *Advances in Neural Information Processing Systems* (I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 1697–1708, 2017.

[33] A. Telenti, S. R. Steinhubl, and E. J. Topol, "Rethinking the medical record," *The Lancet*, vol. 391, no. 10125, p. 1013, 2018.

[34] Department of Veterans Affairs, "Veterans Health Administration: Providing health care for Veterans." https://www.va.gov/health/, 2018 (accessed November 9, 2018).

[35] N. Razavian and D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests," *arXiv Preprint arXiv:1511.07938*, 2015.

[36] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, ACM, 2002.

[37] J. G. Zilly, R. K. Srivastava, J. Koutník, and J. Schmidhuber, "Recurrent highway networks," in *Proceedings of the International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70, pp. 4189–4198, 2017.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] J. Collins, J. Sohl-Dickstein, and D. Sussillo, "Capacity and learnability in recurrent neural networks," *International Conference on Learning Representations*, 2017.

[40] J. Bradbury, S. Merity, C. Xiong, and R. Socher, "Quasi-recurrent neural networks," *International Conference on Learning Representations*, 2017.

[41] T. Lei and Y. Zhang, "Training RNNs as fast as CNNs," *arXiv Preprint arXiv:1709.02755*, 2017.

[42] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv Preprint arXiv:1412.3555*, 2014.

[43] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv Preprint arXiv:1410.5401*, 2014.

[44] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proceedings of the International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), pp. 1842–1850, 2016.

[45] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.

[46] A. Santoro, R. Faulkner, D. Raposo, J. Rae, M. Chrzanowski, T. Weber, D. Wierstra, O. Vinyals, R. Pascanu, and T. Lillicrap, "Relational recurrent neural networks," *arXiv Preprint arXiv:1806.01822*, 2018.

[47] R. Caruana, S. Baluja, and T. Mitchell, "Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation," in *Advances in Neural Information Processing Systems* (M. Mozer, M. Jordan, and T. Petsche, eds.), vol. 9, pp. 959–965, 1996.

[48] J. Wiens, J. Guttag, and E. Horvitz, "Patient risk stratification with time-varying parameters: A multitask learning approach," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2797–2819, 2016.

[49] D. Y. Ding, C. Simpson, S. Pfohl, D. C. Kale, K. Jung, and N. H. Shah, "The effectiveness of multitask learning for phenotyping with electronic health records data," *arXiv Preprint arXiv:1808.03331*, 2018.

[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feed forward neural networks," in *International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9, pp. 249–256, 2010.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

[52] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), pp. 1321–1330, 2017.

[53] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large-Margin Classifiers*, pp. 61–74, MIT Press, 1999.

[54] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[55] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the International Conference on Machine Learning* (L. D. Raedt and S. Wrobel, eds.), pp. 625–632, ACM, 2005.

[56] T. Saito and M. Rehmsmeier, "The precision recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS One*, vol. 10, no. 3, 2015.

[57] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[58] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[59] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems* (I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 6402–6413, 2017.

[60] J. D. Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. W. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. A. Raine, J. C. Hughes, D. A. Sim, C. A. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, pp. 1342–1350, 2018.

[61] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015.

[62] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *European Conference on Computer Vision*, 2014.

[63] J. M. Steppe and K. W. Bauer Jr, "Feature saliency measures," *Computers & Mathematics With Applications*, vol. 33, no. 8, pp. 109–126, 1997.

[64] P. E. Drawz, R. T. Miller, and A. R. Sehgal, "Predicting hospital acquired acute kidney injury - A case controlled study," *Renal Failure*, vol. 30, no. 9, pp. 848–855, 2008.

[65] M. E. Matheny, R. A. Miller, T. A. Ikizler, L. R. Waitman, J. C. Denny, J. S. Schildcrout, R. S. Dittus, and J. F. Peterson, "Development of inpatient risk stratification models of acute kidney injury for use in electronic health records," *Medical Decision Making*, vol. 30, no. 6, pp. 639–650, 2010.

[66] L. G. Forni, T. Dawes, H. Sinclair, E. Cheek, V. Bewick, M. Dennis, and R. Venn, "Identifying the patient at risk of acute kidney injury a predictive scoring system for the development of acute kidney injury in acute medical patients," *Nephron Clinical Practice*, vol. 123, no. 3-4, pp. 143–150, 2013.

[67] R. M. Cronin, J. P. VanHouten, E. D. Siew, S. K. Eden, S. D. Fihn, C. D. Nielson, J. F. Peterson, C. R. Baker, T. A. Ikizler, T. Speroff, and M. E. Matheny, "National Veterans Health Administration inpatient risk stratification models for hospital acquired acute kidney injury," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 1054–1071, 2015.

[68] M. Bedford, P. Stevens, S. Coulton, J. Billings, M. Farr, T. Wheeler, M. Kalli, T. Mottishaw, and C. Farmer, "Development of risk models for the prediction of new or worsening acute kidney injury on or during hospital admission: A cohort and nested study," *Health Service Delivery Research*, vol. 4, no. 6, 2016.

[69] R. J. Kate, R. M. Perez, D. Mazumdar, K. S. Pasupathy, and V. Nilakantan, "Prediction and detection models for acute kidney injury in hospitalized older adults," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, p. 39, 2016.

[70] P. Thottakkara, T. Ozrazgat-Baslanti, B. B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, and A. Bihorac, "Application of machine learning techniques to high dimensional clinical data to forecast postoperative complications," *PLOS One*, vol. 11, no. 5, 2016.

[71] S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny, "Calibration drift in regression and machine learning models for acute kidney injury," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1052–1061, 2017.

[72] L. Hodgson, B. Dimitrov, P. Roderick, R. Venn, and L. G. Forni, "Predicting AKI in emergency admissions: An external validation study of the acute kidney injury prediction score (APS)," *BMJ Open*, vol. 7, no. 3, p. e013511, 2017.

[73] S. J. Weisenthal, H. Liao, P. Ng, and M. S. Zand, "Sum of previous inpatient serum creatinine measurements predicts acute kidney injury in rehospitalized patients," *arXiv Preprint arXiv:1712.01880*, 2017.

[74] L. Adhikari, T. Ozrazgat-Baslanti, P. Thottakkara, A. Ebadi, A. Motaei, P. Rashidi, X. Li, and A. Bihorac, "Improved predictive models for acute kidney injury with IDEAs: Intraoperative data embedded analytics," *arXiv Preprint arXiv:1805.05452*, 2018.

[75] N. Park, E. Kang, M. Park, H. Lee, H.-G. Kang, H.-J. Yoon, and U. Kang, "Predicting acute kidney injury in cancer patients using heterogeneous and irregular data," *PLOS One*, vol. 13, no. 7, 2018.

[76] S. J. Weisenthal, C. Quill, S. Farooq, H. Kautz, and M. S. Zand, "Predicting acute kidney injury at hospital re-entry using high-dimensional electronic health record data," *arXiv Preprint arXiv:1807.09865*, 2018.

[77] Y. Li, L. Yao, C. Mao, A. Srivastava, X. Jiang, and Y. Luo, "Early prediction of acute kidney injury in critical care setting using clinical notes," in *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine*, 2018.

[78] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[79] R. Henao, J. T. Lu, J. E. Lucas, J. Ferranti, and L. Carin, "Electronic health record analysis via deep poisson factor models," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 6422–6453, 2016.

[80] R. Carnevale, S. Mani, A. Ozdas, Y. Chen, Q. Chen, C. Aliferis, H. A. Varol, H. Nian, J. Romano-Keeler, and J.-H. Weitkamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 326–336, 2013.

[81] Z. Hu, G. J Simon, E. G Arsoniadis, Y. Wang, M. R Kwaan, and G. B Melton, "Automated detection of postoperative surgical site infections using supervised methods with electronic health record data," *Studies in Health Technology and Informatics*, vol. 216, pp. 706–10, 08 2015.

[82] L. Wang, W. Zhang, X. He, and H. Zha, "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2018.

[83] E. Choi, A. Schuetz, W. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, p. 112, 2016.

[84] C. Sideris, N. Alshurafa, M. Pourhomayoun, F. Shahmohammadi, L. Samy, and M. Sarrafzadeh, "A data-driven feature extraction framework for predicting the severity of condition of congestive heart failure patients," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2534–2537, Aug 2015.

[85] B. A. Goldstein, T. I. Chang, A. A. Mitani, T. L. Assimes, and W. C. Winkelmayer, "Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic

health records," *Clinical Journal of the American Society of Nephrology*, vol. 9, no. 1, pp. 82–91, 2014.

[86] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proceedings of the 1st Machine Learning for Healthcare Conference* (F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, eds.), vol. 56, pp. 301–318, PMLR, 2016.

[87] Z. Che, S. Purushotham, K. Cho, and D. Sontag, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, p. 6085, 2018.

[88] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Proceedings of the 1st Machine Learning for Healthcare Conference* (F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, eds.), vol. 56, pp. 73–100, PMLR, 2016.

[89] C. Paxton, S. Saria, and A. Niculescu-Mizil, "Developing predictive models using electronic medical records: Challenges and pitfalls," *AMIA Annual Symposium Proceedings*, vol. 2013, pp. 1109–1115, 2013.

[90] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (G. Gordon, D. Dunson, and M. Dudík, eds.), vol. 15, pp. 315–323, PMLR, 2011.

[91] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 30, p. 3, 2013.

[92] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *International Conference on Learning Representations*, 2018.

[93] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *International Conference on Learning Representations*, 2016.

[94] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems* (I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 971–980, 2017.

[95] M. Basirat and P. M. Roth, "The quest for the golden activation function," *arXiv Preprint arXiv:1808.00783*, 2018.