nature

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

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

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

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

Further Supplementary Information
A Clinically Applicable Approach to 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:

- Supplement A shows systematically selected case examples for both correct and incorrect model predictions.
- Supplements B-K provide supplementary methods and results to aid interpretation of the AKI predictions and reproducibility of results. An extensive review of the literature into AKI risk models and machine learning and deep learning for electronic health records is also provided in Supplement F.

Further to this supplementary information, a detailed protocol paper entitled "*Developing Deep Learning Continuous Risk Models for Early Adverse Event Prediction in Electronic Health Records: an AKI Case Study*" has been made available though *Protocol Exchange*.

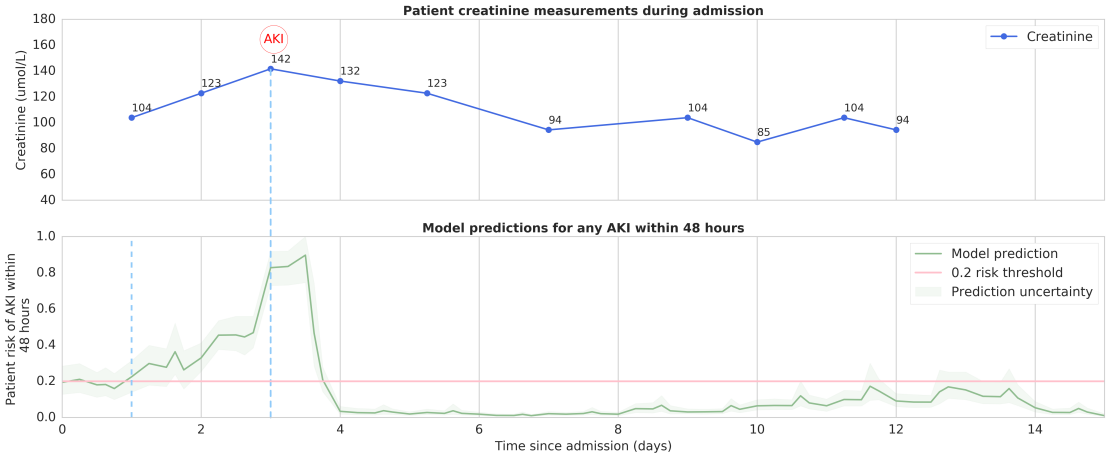
A. Success and failure cases

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.

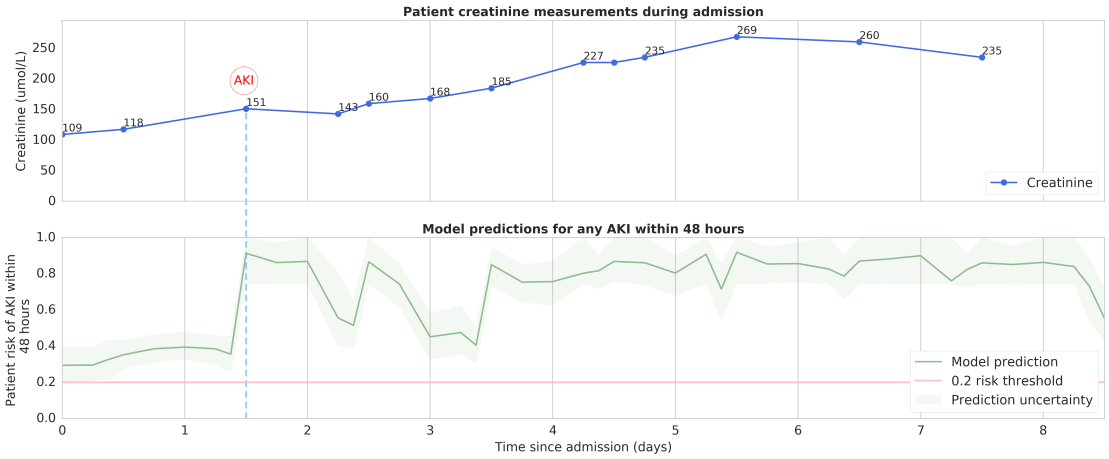
To avoid demonstrating the performance of the model by 'cherry picking' a single example, we present an additional set of five systematically selected success and failure cases of the predictive model. In each of these examples, the first plot shows the creatinine measurements throughout the admission 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.

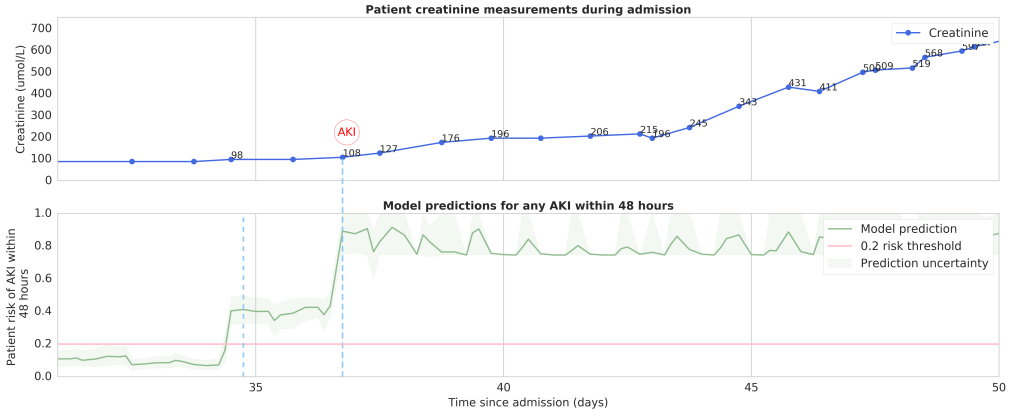
A.1. Success case examples



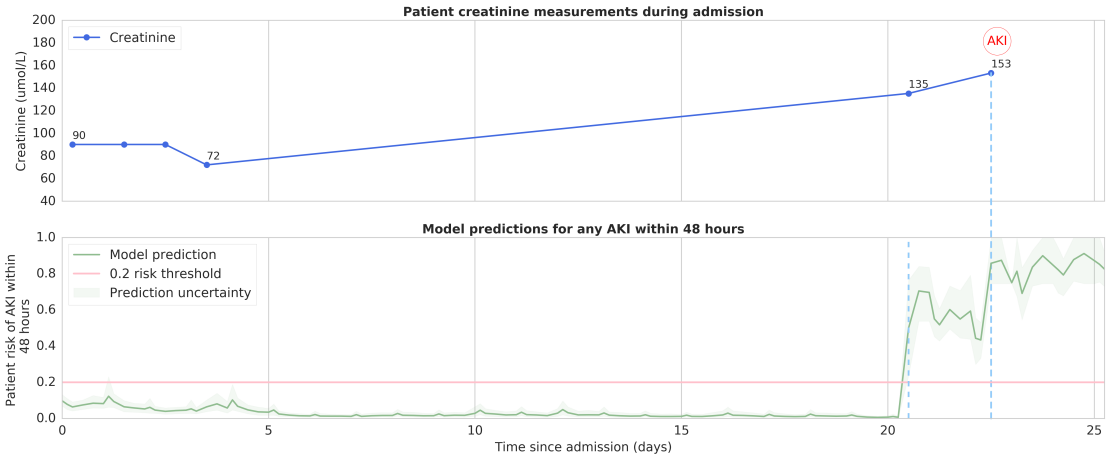
Supplementary Figure 1 | Visual representation of a 15 day surgical admission for a 77 year old male patient with a history of congestive heart failure. The patient developed AKI 3 days after admission, with accompanying evidence of sepsis. The model correctly predicts the patient is at risk 48 hours before the AKI is detected according to KDIGO criteria.



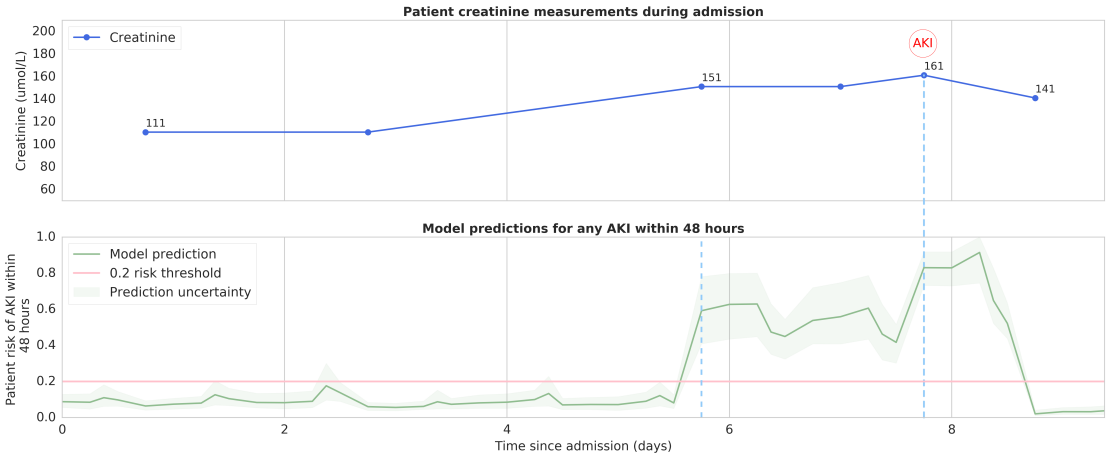
Supplementary Figure 2 | Visual representation of a 9 day intensive care admission for a 57 year old male with a history of diabetes. The first onset of AKI occurs during the second day of admission; from the beginning of the admission the model predicts the risk at above the 0.2 threshold. Ultimately the patient went on to develop chronic kidney disease after discharge.



Supplementary Figure 3 | 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.

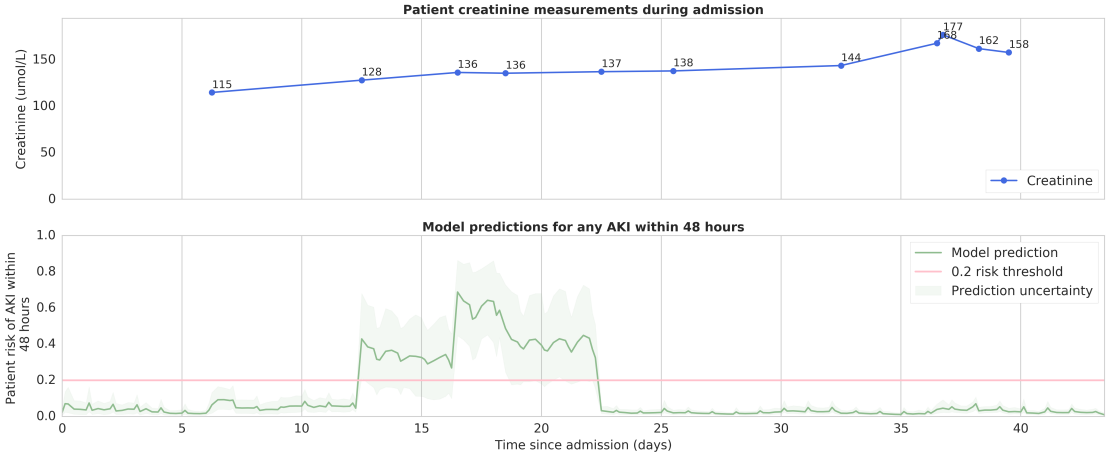


Supplementary Figure 4 | Visual representation of an admission under the medical team of a 64 year old male with a history of CKD and congestive heart failure. After a long period without blood measurements, the patient developed an AKI on the 22nd day of admission, which was correctly anticipated by the model.

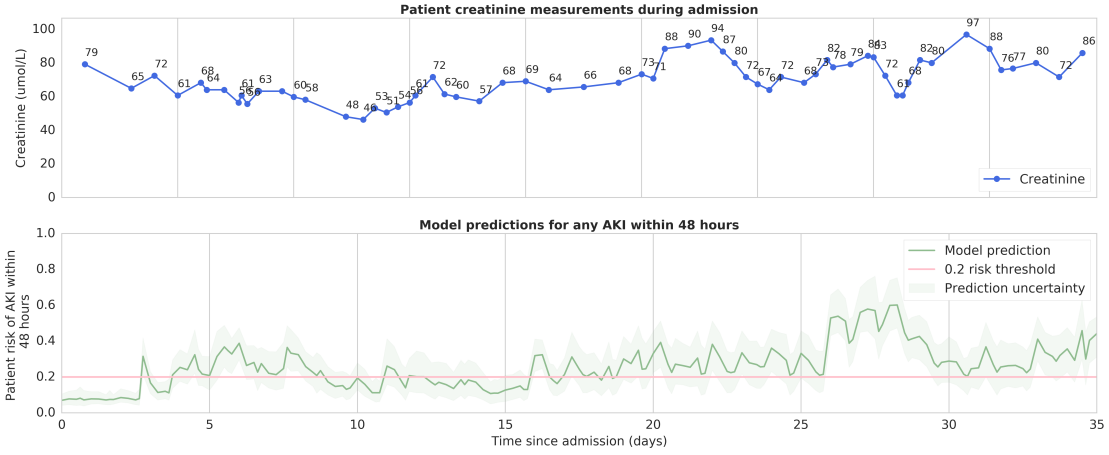


Supplementary Figure 5 | A visual representation of a 10 day medical admission of a 60 year old male with a history of congestive heart failure. The model correctly predicts the gradual increase of creatinine being labelled as AKI by KDIGO criteria.

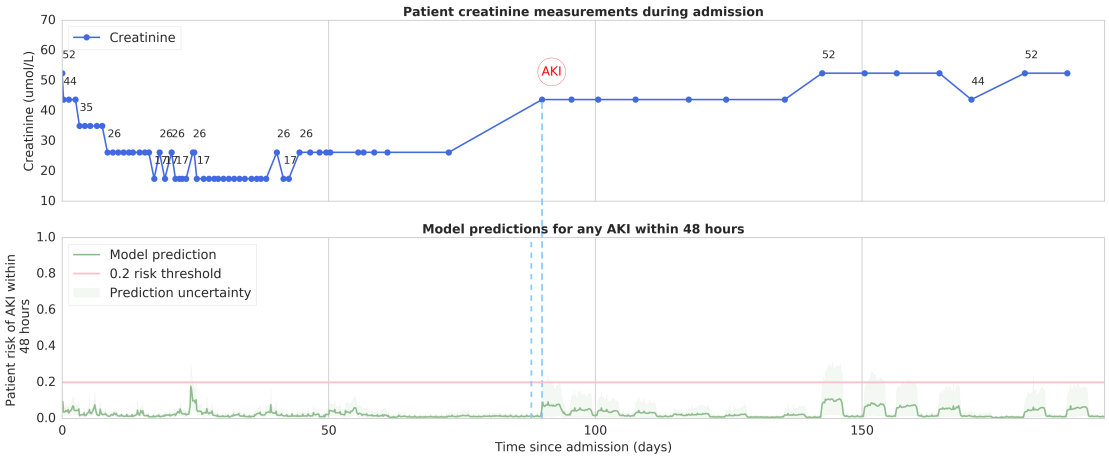
A.2. Failure case examples



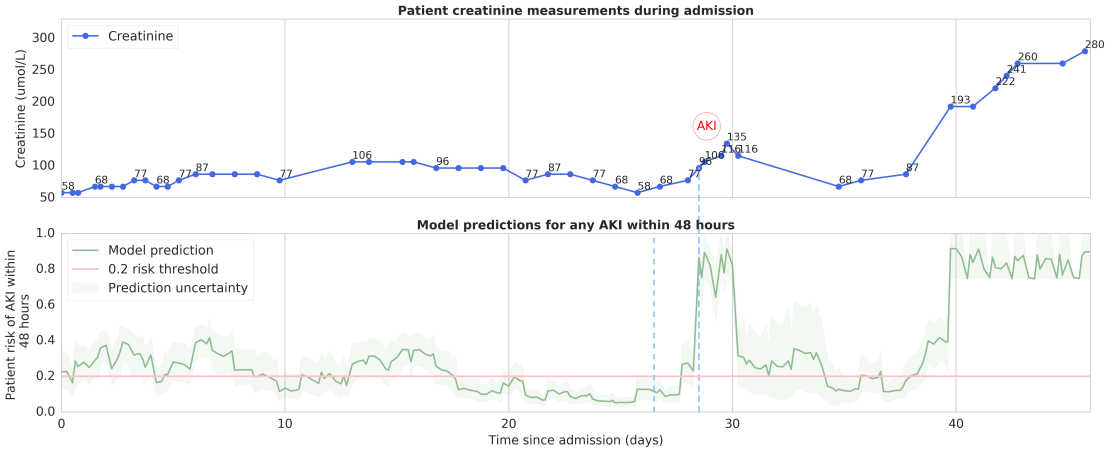
Supplementary Figure 6 | A 59 year old male with a history of CKD, admitted under the medical team with evidence of sepsis and transferred to the intensive care unit 2 days after admission. Despite infrequent creatinine measurements in the patient records, e-GFR is consistently measured, suggesting information 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.



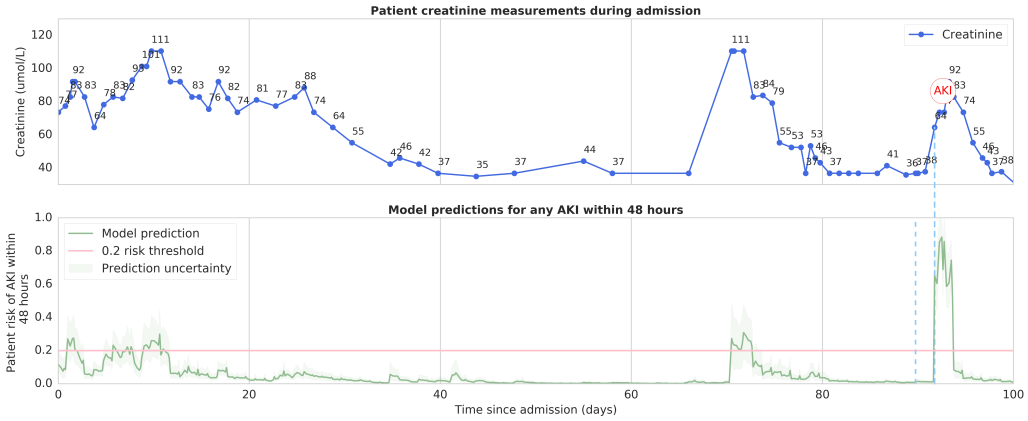
Supplementary Figure 7 | A 57 year old male with multiple previous AKI episodes in previous admissions, admitted here with evidence of infection. Despite a long 35 day admission with frequently raised inflammatory markers the patients renal function remained stable; the model provides raised risk scores throughout this admission.



Supplementary Figure 8 | 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 9 | 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 10 | 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.

B. 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 1 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.

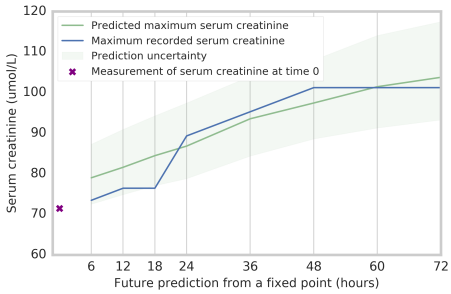
Supplementary Table 2 shows the accuracy of the model in predicting the trajectory of the selected laboratory values 48 hours ahead of time. Supplementary Figure 11 shows an example of these predictions for a given admission.

Supplementary Table 1 | 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.

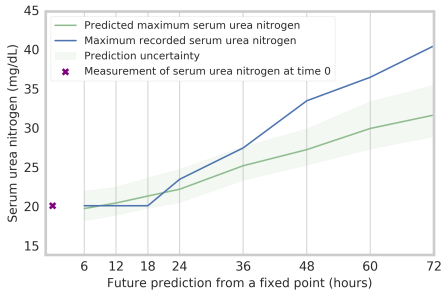
Laboratory test	Units	Subgroup	Number samples (000s)	Subgroup mean	Subgroup standard deviation	Absolute error (SRU)	Relative error (%) (SRU)	Relative error (%) (LR)
Serum urea nitrogen	mg/dL	Population	2912.4	21.6	14.5	3.4	18.7 [18.6, 18.7]	89.7 [69.0, 101.6]
		AKI in 48 hours	188.9	36.4	19.8	7.6	21.3	
		>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 2 | 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.

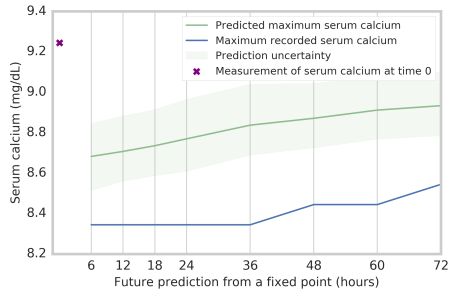
	% predictions correctly predicting an increase in value in 48 hours	
Laboratory test	Cases where the value is increasing	Cases where the value is increasing by an amount more than the median
Serum urea nitrogen	83.7%	90.8%
Serum creatinine	83.6%	86.3%
Serum potassium	85.2%	90.5%
Serum sodium	79.4%	88.5%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



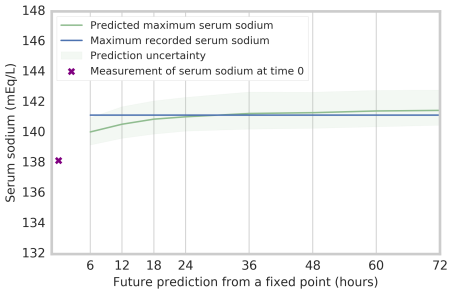
(a) Serum creatinine



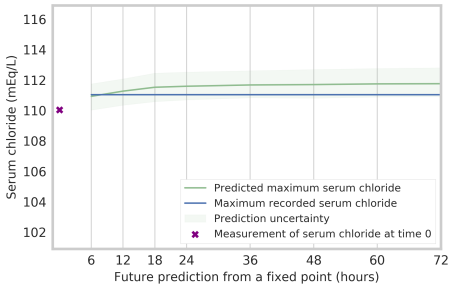
(b) Serum urea nitrogen



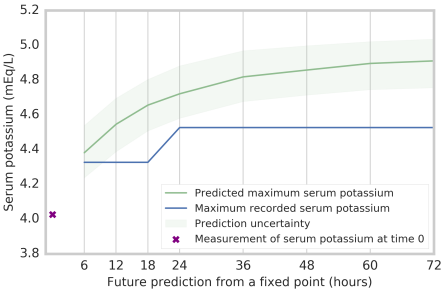
(c) Serum calcium



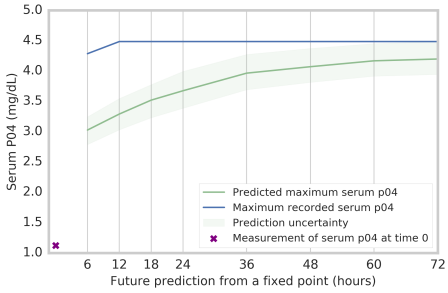
(d) Serum sodium



(e) Serum chloride



(f) Serum potassium



(g) Serum phosphate

Supplementary Figure 11 | 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.

C. 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 [1]. 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 3. 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 [2]. 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 3 | 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 creatinine baselines shown in Supplementary Table 3 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 3. 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.

D. 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 4 | 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 n=200 bootstrap samples per model. 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]

E. 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.

F. Literature review

F.1. AKI risk models

Supplementary Table 5 | 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 [3]	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 [4]	U.S.	1	Adults with admissions of ≥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 [5]	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 [6]	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 [7]	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 [8]	U.S.	15	Patients ≥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 [9]	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 [10]	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 [11]	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 [12]	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 [13]	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 [14]	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 [15]	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 [16]	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 [17]	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 [18]	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 [19]	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 [20]	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 [21]	U.S.	1	ICU patients	∼40,000	-	KDIGO	24h after admission	Y	Convolutional Neural Network	77.9%	Precision: 40.7% Recall: 65.4%
Pan 2019 [22]	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

F.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 [23]. We provide a systematic overview of these achievements in Supplementary Table 6. Machine learning models have shown promise when used for predicting mortality [24–26], sepsis [10, 27, 28], post-operative complications [17, 29], readmission risk [30], for providing treatment recommendations [31], modelling treatment response [32, 33], detecting early signs of heart failure [34–36] and in planning for palliative care [37]. Most of the deep learning approaches involve improvements in representation learning [38] or apply recurrent neural networks (RNN) [24, 27, 34, 39–41] or convolutional models [30, 42–44].

Despite these recent advances, building robust clinically applicable risk models from routinely collected EHR data remains a challenge [45]. 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 6 | 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 [24]	10,980	-	87	Mortality, cystic fibrosis, comorbidities	LSTM + additional layers
Rajkomar 2018 [25]	114,003	216,221	all available data	Mortality, readmission, long length of stay, discharge diagnosis	LSTM, TANN, boosted decision stumps
Futoma 2017 [27]	-	49,312	77	Sepsis	GP + LSTM
Nguyen 2016 [30]	∼300,000	590,546	diagnoses, procedures	Readmission	CNN
Wang 2018 [31]	∼43,000	22,865	-	Treatment optimisation	SRL-RNN
Avati 2017 [37]	221,284	-	13,654	(3-12 month) Mortality	MLP
Miotto 2016 [38]	∼700,000	-	41,072	Disease prediction	stacked denoising AEs
Lipton 2017 [39]		10,401	13	Diagnosis classification	LSTM
Choi 2016 [40]	263,706	-	1,778	Predicting properties of subsequent visits	GRU
Choi 2016 [34]	32,787	-	diagnoses, procedures, medication	Heart failure detection	GRU
Che 2016 [41]	-	58,000	99	Mortality, diagnosis category	GRU-D
Razavian 2016 [42]	∼298,000	-	44	CKD progression	CNN, LSTM
Cheng 2016 [44]	319,650	-	diagnoses	Congestive heart failure, chronic obstructive pulmonary syndrome	CNN
Komorowski 2018 [46]	96,156	-	48	Sepsis treatment	MDP
Henao 2016 [26]	240,000	4,400,000	24,567	Mortality and morbidity	Deep Poisson factor models
Soleimani 2017 [32]	67	-	5	Dialysis treatment response	Gaussian processes
Schulam 2017 [33]	428	-	4	Dialysis treatment response	Gaussian processes
Alaa 2016 [47]	6,313	-	12	Risk of adverse events	Hierarchical latent class model and Gaussian processes
Thottakkara 2016 [10]	50,318	-	285	Post-operative AKI and sepsis	Naive Bayes and SVM
Bihorac 2018 [17]	51,457	-	-	Post-operative complications	Generalised additive model
Perotte 2015 [48]	2,908	-	106	CKD progression	Kalman filter and Cox proportional hazards
Hu 2015 [29]	6,258	-	demographics, diagnoses, orders, labs, vitals, medications	Surgical site infections	Logistic regression
Sideris 2015 [35]	3,041	-	demographics, diagnoses, labs	Heart failure	SVM + clustering
Goldstein 2014 [36]	1,718	-	72	Sudden cardiac death	Random forests
Mani 2014 [28]	299	1826	811	Neonatal sepsis	Random forests SVM CART Logistic regression
Henry 2015 [49]	16,234	-	54	Sepsis	Cox proportional hazards model

G. 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 7 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 8 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 7. 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 7.

Supplementary Table 7 | 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 8 | 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	<1e-6	[3.93, 4.03]
Age group 20 to 30	0.64	0.05 <1e-6	[0.54, 0.75]	
Age group 30 to 40	0.30	0.03	<1e-6	[0.24, 0.36]
Age group 40 to 50	0.26	0.02	<1e-6	[0.23, 0.30]
Age group 60 to 70	-0.06	0.01	<1e-6	[-0.07, -0.04]
Age group 70 to 80	0.01	0.01	0.196	[-0.01, 0.03]
Age group 80 to 90	0.19	0.01	<1e-6	[0.17, 0.22]
Ethnicity: Black	-0.14	0.01	<1e-6	[-0.15, -0.13]
Gender: Female	0.15	0.02	<1e-6	[0.12, 0.19]
Patients with CKD	-0.62	0.01	<1e-6	[-0.64, -0.61]
CKD stage 1	0.16	0.01	<1e-6	[0.14, 0.18]
CKD stage 2	-0.08	0.01	<1e-6	[-0.11, -0.06]
CKD stage 3a	-0.23	0.01	<1e-6	[-0.25, -0.21]
CKD stage 3b	-0.56	0.01	<1e-6	[-0.59, -0.54]
CKD stage 4	-0.95	0.01	<1e-6	[-0.98, -0.93]
CKD stage 5	-1.09	0.03	<1e-6	[-1.14, -1.05]
Medical admissions	-0.16	0.01	<1e-6	[-0.17, -0.15]
Surgery admissions	-0.19	0.01	<1e-6	[-0.20, -0.17]
ICU transfers	-0.31	0.01	<1e-6	[-0.33, -0.30]
ER visits	0.09	0.01	<1e-6	[0.08, 0.11]
Diabetic patients	-0.11	0.01	<1e-6	[-0.12, -0.09]
Death within 30 days of admission	-0.17	0.02	<1e-6	[-0.20, -0.14]
Death within 7 days of admission	-0.14	0.02	<1e-6	[-0.17, -0.10]
Haemoglobin <80g/L	-0.23	0.01	<1e-6	[-0.25, -0.22]
Haemoglobin <80g/L in first 2 days	0.011	0.01	0.110	[-0.00, 0.04]
WCC >12 or <3.5 x10^9/L	-0.01	0.01	0.297	[-0.03, 0.01]
WCC >12 or <3.5 x10^9/L in first 2 days	-0.15	0.01	<1e-6	[-0.17, -0.14]
Admission duration > 7 days	0.11	0.01	<1e-6	[0.10, 0.13]
Post IV contrast administration	-0.04	0.01	<1e-6	[-0.05, -0.03]
Post IV saline administration	-0.23	0.02	<1e-6	[-0.27, -0.20]
Old information aggregation only	0.30	0.01	<1e-6	[0.29, 0.31]
Admission with at least 1 AKI	-0.93	0.02	<1e-6	[-0.97, -0.89]

H. 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 9). 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 9 | 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]	< 1e-6
All entries	$\leq$ 24 hours	206.3	[205.2, 207.5]	168.8	[168.3, 169.3]	< 1e-6
Serum creatinine	$\leq$ 12 hours	0.83	[0.82, 0.84]	0.64	[0.64, 0.65]	< 1e-6
Serum creatinine	$\leq$ 24 hours	1.25	[1.24, 1.26]	1.00	[1.00, 1.01]	< 1e-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 n=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. 95% bootstrap pivot confidence intervals are calculated using n=200 bootstrap samples.

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. 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 12.

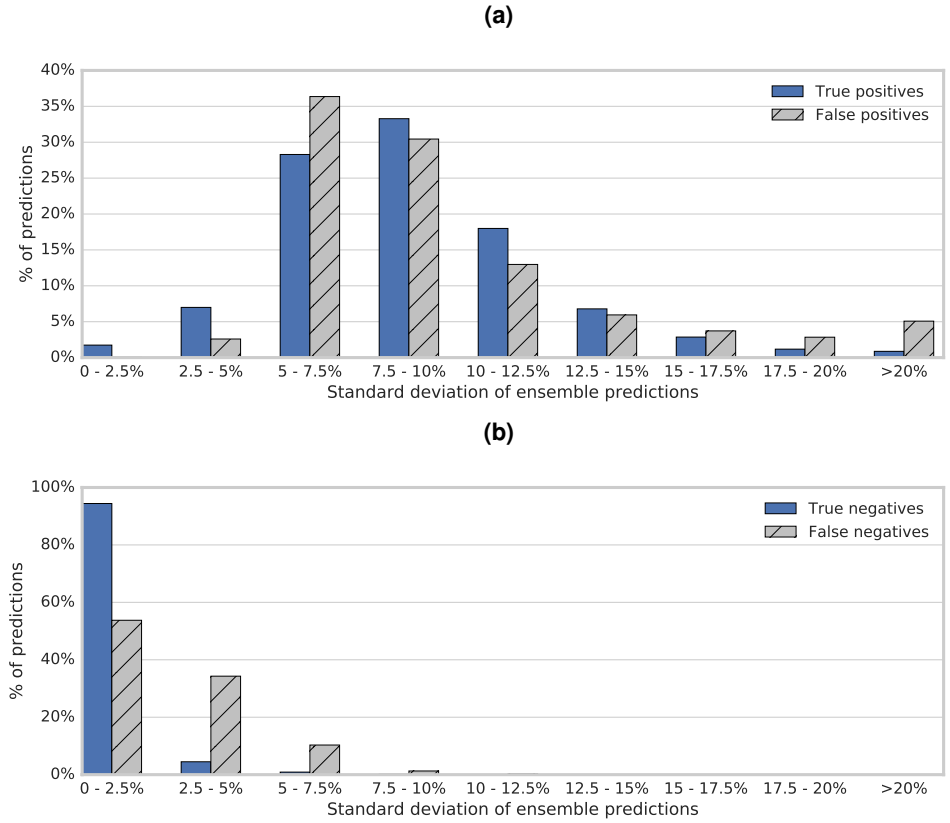
Supplementary Table 12 | Hyperparameter combinations evaluated in the experiments

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 [50], Leaky ReLu [51], Swish [52], ELU [53], SELU [54], ELiSH [55], Hard ELiSH [55], 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

K. 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 12 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 < 1e-6) 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 12 | 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 predictions), 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 predictions) as evidenced by a 2-sided T-test (p-value < 0.01). Figure **b** shows that for true negative predictions (N=8,907,932 predictions), 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 predictions) as evidenced by a 2-sided T-test (p-value < 1e-6).

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

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

[54] 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.

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