Aus der Klinik für kardiovaskuläre Chirurgie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Deep-learning basierte Echtzeit-Vorhersage von akutem Nierenversagen nach kardiochirurgischen Eingriffen

Deep-learning based real-time prediction of acute kidney injury after cardiac surgery

zur Erlangung des akademischen Grades
Medical Doctor - Doctor of Philosophy (MD/PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Nina Rank

Datum der Promotion: 03.03.2023

Table of Contents

Abbreviations

AI	Artificial intelligence
AKI	Acute kidney injury
AUC	Area under the curve
CABG	Coronary artery bypass grafting
CK	Creatine kinase
CKD	Chronic kidney disease
CI	Confidence interval
CPB	Cardiopulmonary bypass
CRP	C-reactive protein
CVP	Central venous pressure
ECMO	Extracorporeal membrane oxygenation system
EHR	Electronic health record
FPR	False-positive rate
GBM	Gradient boosted machine
GRU	Gated recurrent unit
ICU	Intensive care unit
KDIGO	Kidney Disease: Improving Global Outcomes
LDH	Lactate dehydrogenase
LSTM	Long-short-term memory
ML	Machine learning
MSE	Mean squared error
MSE_{pat}	Mean squared error of individual patient
NN	Neural network
NPV	Negative predictive value
PPV	Positive predictive value
PR_AUC	Precision-recall area under the curve
RNN	Recurrent neural network
TAVI	Transcatheter aortic valve implantation
TMVI	Transcatheter mitral valve implantation

Abstract (deutsch)

Die zunehmende Digitalisierung medizinischer Daten und die Fortschritte im Bereich der künstlichen Intelligenz ermöglichen es, die enorme Menge an Daten, die während eines Krankenhausaufenthalts gesammelt wird, auf viel komplexere Weise zu nutzen, als es bislang der Fall war. In der im Rahmen der Promotion durchgeführten Studie wurde dieser Ansatz für die Echtzeit-Vorhersage von postoperativem akutem Nierenversagen (ANV) verfolgt – eine der häufigsten Komplikationen nach kardiothorakalen Eingriffen. Anhand von 96 Parametern, die standardmäßig während eines Krankenhausaufenthalts aufgezeichnet werden, wurde ein rekurrentes neuronales Netz (RNN) entwickelt, das ANV innerhalb der ersten sieben postoperativen Tage vorhersagen kann. Das Modell wurde mit Daten aus n = 2224 Aufnahmen trainiert, welche aus n = 15.564 klinischen Fällen in einem Krankenhaus der tertiären Versorgung für kardiothorakale Chirurgie zusammengestellt wurden. Die Leistung des RNN wurde anhand eines unabhängigen Testsets aus n = 350 klinischen Fällen bewertet, und es wurde eine *area under the curve* (AUC) (95 % Konfidenzintervall) von 0,893 (0,862 - 0,924) ermittelt. Zusätzlich wurde ein direkter Vergleich der Vorhersagegüte zwischen dem RNN und erfahrenen ÄrztInnen durchgeführt. Das RNN übertraf die ÄrztInnen in Bezug auf alle ermittelten statistischen Messwerte (z.B. AUC = 0,901 vs. 0,745, p < 0,001). Im Gegensatz zu den Vorhersagen der ÄrztInnen, die das Risiko der Entwicklung eines ANV generell unterschätzten, zeigte das RNN eine gute Kalibrierung. Die Integration eines solchen Modells in bestehende elektronische Patientendatensysteme könnte durch frühzeitige Vorhersage von ANV ermöglichen, präventive Maßnahmen rechtzeitig zu ergreifen, um Komplikationen zu verhindern. Es könnte als Echtzeit-Überwachungssystem eingesetzt werden und die Entscheidungsprozesse der ÄrztInnen unterstützen. Bei der Verwendung eines solchen Systems sind neben seiner Vorhersagegüte aber auch ethische und rechtliche Aspekte zu berücksichtigen, die den Datenschutz, die Modellentwicklung und den klinischen Einsatz betreffen, und die in dieser Arbeit ebenfalls erörtert werden.

Abstract (english)

The increasing digitisation of medical data and advances in artificial intelligence have enabled us to use the tremendous amount of data that is recorded during a hospital stay in a much more sophisticated way than is currently the case. In the study undertaken and published in the context of this doctoral project, this approach was taken for predicting postoperative acute kidney injury (AKI) – one of the most common and severe complications after cardiothoracic interventions. Using 96 parameters, standardly recorded during a hospital stay, a recurrent neural network (RNN) was developed that predicted AKI within the first seven postoperative days. The training of the model was based on n = 2224 admissions gathered from n = 15,564 admissions at a tertiary care hospital for cardiothoracic surgery. The performance of the model was assessed using an independent test set of n = 350 clinical cases and an *area under the curve* (AUC) (95% confidence interval) of 0.893 (0.862 - 0.924) was obtained. Additionally, a head-to-head comparison of the RNN against experienced physicians was conducted. The RNN exceeded the physicians in terms of all determined statistical measures (*e.g.*, AUC = 0.901 vs 0.745, $p < 0.001$). In contrast to the predictions of physicians, who generally underrated the risk of developing AKI, the RNN showed good calibration. The integration of such a model into existing digital medical record systems could allow preventive steps to be taken in time to prevent complications by predicting AKI well before its onset. It could be used as a real-time surveillance system and support physicians' decision-making process. However, when using such a technique, there are several ethical aspects to be considered concerning data protection, model development, and clinical deployment, which are also discussed in this work.

1. Introduction

Patients undergoing cardiac surgery are highly prone to develop various postoperative complications ranging from heart failure, postoperative bleeding, stroke, sepsis, complications of the central nervous system, the kidney and the respiratory system (Ball et al., 2016). These complications significantly impact patients' outcomes in the postoperative period as well as their long-term survival (Pahwa et al., 2021). Early identification of patients at high risk could help to prevent or mitigate such complications by early intervention. The study carried out and published in the context of this doctoral project describes the successful development and evaluation of an innovative machine learning prediction tool for postoperative acute renal failure.

In Chapter 1 of this synopsis report, the definition of acute kidney injury (AKI), its relation to cardiothoracic surgery, its impact on the economy and patients' health and the opportunities that increasing digitisation of medical data opens up in the prediction of AKI are presented. Chapter 2 is dedicated to the methodology, whereas the results obtained are presented in Chapter 3. Under Chapter 4, possible clinical applications, further research questions and limitations of the study are discussed. Chapter 4 also includes a thorough assessment of ethical and legal considerations as regards the possible future clinical deployment of ML-based applications. Concluding remarks form the final Chapter 5 of this synopsis. Partial findings of the present work, especially those related to the current state of research, the methods and the results, were published in *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139.

1.1 The importance of predicting acute kidney injury after cardiac surgery

1.1.1 Definition of acute kidney injury

AKI is characterized by a sudden deterioration of renal function that occurs within hours or days and is in principle reversible. Different AKI stages are distinguished according to KDIGO guidelines as described in Table 1 (Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012):

Table 1. Stages of acute kidney injury according to KDIGO (Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012).

Stage	Creatinine Criteria	Urine Output Criteria
1	Increase of ≥ 0.3 mg/dl (≥ 26.5 μmol/l) within 48 hours OR Increase to 1.5 - 1.9 times baseline within 7 days	< 0.5 ml/kg/h for 6 - 12 hours
2	Increase to 2.0 - 2.9 times baseline	< 0.5 ml/kg/h for ≥ 12 hours
3	Increase to 3.0 times baseline OR Increase in serum creatinine to ≥ 4.0 mg/dl (≥ 353.6 μmol/l) OR Initiation of renal replacement therapy OR In patients < 18 years, decrease in estimated glomerular filtration rate (eGFR) to < 35 ml/min per 1.73 m^2	< 0.3 ml/kg/h for ≥ 24 hours OR Anuria for ≥ 12 hours

1.1.2 Acute kidney injury after cardiac surgery

AKI is a common and severe complication after cardiothoracic surgery and is still not entirely understood (Wang & Bellomo, 2017). Multiple risk factors for AKI development are described and can be divided into different groups. The first group comprises general patient-related risk factors like high age, female gender, reduced left ventricular function, diabetes and chronic kidney disease (Rosner & Okusa, 2006). The second group composes cardiothoracic surgery-associated risk factors such as type of surgery, usage, duration and flow characteristics (pulsatile vs non-pulsatile) of cardio-pulmonary bypass, aortic cross-clamp time or hemodilution. In addition, there are several postoperative factors that increase the risk of AKI, such as low cardiac output, hypotension, inflammation and oxidative stress due to surgical injury, sepsis, atheroembolism and usage of nephrotoxins (Wang & Bellomo, 2017).

1.1.3 Impact of acute kidney injury

It has been shown that AKI is an independent risk factor for short- and long-term mortality (Glenn M. Chertow et al., 1998; C. E. Hobson et al., 2009; Mandelbaum et al., 2011; Ympa et al., 2005). Dasta & Kane-Gill reported that AKI in hospitalized patients increases the mortality rate 4- to 10-fold (Dasta & Kane-Gill, 2019). In addition, AKI leads to prolonged length of hospital stays and elevated treatment costs (C. Hobson et al., 2015; Silver et al., 2017; Silver & Chertow, 2017). Studies revealed that the costs for patients with AKI in intensive care units (ICUs) are about twice as high as those of patients without AKI and that the hospitalizations costs of AKI exceed those of gastrointestinal bleeding and myocardial infarction (Dasta & Kane-Gill, 2019; Silver et al., 2017). Thus, postoperative AKI leads to a considerable financial burden on the healthcare system.

In patients in whom AKI develops into end-stage renal failure with the need of life-long renal replacement therapy, further socioeconomic consequences must also be considered. It has been shown that dialysis hampers employment status (Nakayama et al., 2015) and considerably reduces the quality of life of the affected patients (Dąbrowska-Bender et al., 2018).

1.1.4 Prediction of acute kidney injury

It is therefore desirable to prevent kidney failure by early measures whenever possible. A study by Balasubramanian et al. showed that early nephrologist involvement in patients with AKI stage 1 could prevent further deterioration of kidney function (Balasubramanian et al., 2011). In contrast, delayed nephrologist consultation was accompanied by elevated mortality rates and dialysis dependence in critically ill patients with AKI (Costa e Silva et al., 2013). Meersch et al. revealed that the risk of cardiac surgery-associated AKI can be reduced by administration of an immediate postoperative “KDIGO care bundle” including *“optimization of volume status and hemodynamics, avoidance of nephrotoxic drugs, and preventing hyperglycemia in high risk patients”* (Meersch et al., 2017).

Due to the multifactorial etiology of AKI and the complex interactions of risk factors, the prediction of AKI remains, however, a difficult task. Serum creatinine is an insufficient marker for the early identification of high-risk patients since it only increases when the kidney function is already considerably impaired (Murty et al., 2013). Particularly in elderly patients, who often have diminished muscle mass and subsequently lower serum creatinine levels, consideration of serum creatinine levels alone leads to the underdiagnosis of renal failure (Swedko et al., 2003).

Several clinical risk scores for AKI are available (Aronson et al., 2007; G. M. Chertow et al., 1997; Huen & Parikh, 2012; Mehta et al., 2006; Palomba et al., 2007; Thakar et al., 2005; Wijeysundera et al., 2007). However, there is no consensus recommendation as to which one to use. Most of these classical risk scores only implement static variables like clinical history, demographics and surgery type and are thus not able to adapt to sudden changes in patients' states. Additionally, they usually demand additional workload for clinical staff as the data collection is not automated.

Several novel biomarkers to identify AKI have been developed and evaluated (Bennett et al., 2008; Burke-Gaffney et al., 2014; Haase et al., 2009; Jayakumar et al., 2013; Krawczeski et al., 2011; McIlroy et al., 2010; Mishra et al., 2005; Parikh et al., 2011; Ramesh et al., 2010), but their benefit over clinical assessment remains uncertain. Moreover, most of them are widely unavailable and, partly due to unclear cost-effectiveness, not part of routine diagnostics (Wang & Bellomo, 2017).

The recent progression of digitalisation in the medical sector has created the opportunity to use medical information now much more sophisticated by capturing underlying information in the data that would otherwise be overlooked. The tremendous amount of data that accumulates during a hospital stay is, however, too overwhelming for clinical staff to effectively be processed in limited time and in the often stressful environment of ICUs (Donchin & Seagull, 2002; Halford et al., 2005). Latest developments in artificial intelligence (AI) could potentially overcome this problem by automatically analysing high dimensional data, predicting future outcomes and thus providing decision support for physicians.

1.2 Current state of research

The application of ML to complex medical problems like AKI is not new and has already achieved auspicious results (Rank et al., 2020). In 2016, Thottakkara et al. used different ML algorithms to predict postoperative AKI and yielded *areas under the curve* (AUCs) between 0.797 and 0.858 in their internal validation set (Thottakkara et al., 2016). Bihorac et al. applied ML to evaluate the risk of multiple postoperative complications and observed an AUC of 0.80 (0.79-0.80) for AKI (Bihorac et al., 2019). In both studies, however, only static, mostly preoperative parameters were used for prediction.

In 2016, Koyner et al. conducted a multi-center ward-based study and built a discrete-time survival model which yielded an AUC (95% CI) of 0.76 (0.76-0.77) for AKI of stage ≥ 2 (J. L. Koyner et al., 2016). Another study based on electronic health record (EHR) data by Koyner et al. followed in 2018 in which the research group obtained an AUC (95% CI) of 0.90 (0.90–0.90) at forecasting AKI stage 2 within the following 24 hours and 0.87 (0.87–0.87) within the following 48 hours (Jay L. Koyner et al., 2018). Cheng et al. used ML to predict AKI over multiple time spans and yielded an AUC of 0.765 for the prediction at one day before AKI onset (Cheng et al., 2017). These studies, however, did not incorporate the urine criterion of AKI (see Table 1), which may result in false-negative labelling of the AKI cases. Moreover, only patients with a serum creatinine of < 3mg/dl (Koyner et al., 2018) or normal serum creatinine and a GFR of ≥ 60ml/min/1.73m^2 (Cheng et al., 2017) at admission were included in the studies. Especially for patients with already impeded kidney function, however, close postoperative monitoring and AKI risk prediction should be desired.

Based on EHR data, Mohamadlou et al. built an ML model for AKI detection and AKI prediction 12 to 72 hours before AKI onset, for which they yielded AUCs from 0.872 (at onset) - 0.728 (72h before onset) (Mohamadlou et al., 2018). A particularly large study was reported in 2019 by Tomašev et al., in which the research group developed a recurrent neural net (RNN) for continuous prediction of AKI (Tomašev et al., 2019). They achieved an AUC up to 0.971 24h before onset. However, these studies also did not integrate the urine output criterion of AKI. Moreover, Tomašev et al. only included patients with at least one year of available medical history in the EHR system. Additionally, they incorporated aggregated historical medical data collected over up to five years. In a real setting, however, patients are not always known prior to admission, and the performance of the algorithm on patients without this information remains unclear. The model developed in this doctoral project, however, only incorporated time-series data that was recorded after or directly at admission but no historical information.

An RNN for the prediction of AKI requiring dialysis, mortality and postoperative bleeding after cardiac surgery within the 24 postoperative hours was also developed by Meyer et al., based on a stream of peri- and postoperative routinely collected data (Meyer et al., 2018). Their model performed well (positive predictive value of 0.87 and sensitivity of 0.94 for AKI) and surpassed classical clinical risk assessment scores.

Using ML to predict AKI after cardiac surgery continues to be a highly topical issue. Since the publication of the study underlying this doctoral project in 2020, Penny-Dimri et al. in 2021 compared the performance of four ML algorithms (logistic regression, K-Nearest-Neighbours, gradient boosted machine (GBM) and neural networks (NN)) with that of two established scores used to predict cardiac-surgery associated AKI (Penny-Dimri et al., 2021). Logistic regression, GBM and NN outperformed the latter. In addition, they managed to extract patient-level risk profiles for their predictions from GBM and NN. This information is particularly valuable in clinical practice and drives personalised medicine forward.

Despite these very promising results, to date, no other study apart from the one carried out for this doctoral project exists that compares the performance of an ML algorithm with that of experienced clinicians in predicting postoperative AKI on longitudinal data streams of real-world hospital cases.

1.3 Significance of this doctoral project for the prediction of acute kidney injury

The study underlying this doctoral project aimed to first develop a machine learning (ML) algorithm that predicts AKI after cardiothoracic surgery based on standardly recorded parameters. More specifically, the algorithm was intended to allow real-time predictions, meaning that it should estimate the risk of developing AKI at any point in time during a patient’s observation period and not only give a static preoperative prediction. Such a system requires constant adaptation to changes in patients' state of health. The model was intended to be designed in a way as to allow a potential integration into EHR systems, which could enable real-time monitoring of patients, early detection of imminent AKI and, thus, initiation of preventive measures.

The algorithm was designed to forecast AKI up to the first seven postoperative days. This time span is much longer than that of the described studies. Usually, closer events can be predicted more easily than events in the far future. However, action should be taken as soon as renal failure is imminent and not only when the kidneys are already - and possibly irreversibly - damaged. Therefore a wider prediction horizon is desirable as studies revealed that timely intervention could avert severe AKI (Balasubramanian et al., 2011; Meersch et al., 2017).

A further goal of the study was to compare the model’s performance against that of experienced physicians. An important criterion for the introduction of ML algorithms into the clinical routine is that their performance should not be substantially worse than that of human physicians. Thus, this head-to-head comparison was designed as a non-inferiority experiment.

2. Methodology

2.1 Ethical approval

Ethical approval of the study was obtained from the institutional data protection officer and ethics committee of Charité – Universitätsmedizin Berlin (EA2/180/17). This approval comprised the acquisition of data on implied consent. Only retrospective medical information was used, and the patients did not actively participate in the study. The Institutional Review Board of Charité - Universitätsmedizin Berlin waived the requirement of informed consent of the participating physicians as the data collection was anonymized. The description of the model design and its evaluation are broadly in line with the guideline in the TRIPOD statement (Collins et al., 2015).

2.2 Study population and data retrieval

The data employed in the study underlying this doctoral project were retrieved from the EHR system of Deutsches Herzzentrum Berlin and were generated between 10/2012 and 02/2018. The patient selection procedure is illustrated in Fig. 1 (Rank et al, 2020). Initially, all adult patients having received cardiothoracic surgery in this period were included. After exclusion of 2586 admissions (exclusion criteria: no creatinine/urine flow values available, hemodialysis before the end of the operation, baseline creatinine ≥ 4.0mg/dl), 1308 cases were identified with AKI stage 2 or 3 within seven days after surgery.

The transfer of a patient to the ICU/recovery room denoted the starting point of the observation time of the respective patient. The respective endpoint was defined as soon as one of the following criteria apply:

- at least one KDIGO criterion for AKI stage 2 or 3 was fulfilled
- the patient was discharged
- seven days after the end of the surgery were completed

A balanced data set was then created by assigning each AKI-case a non-AKI-control. The pairs were matched by observation length. This data set was then randomly split into training (2224 admissions/2180 patients, 85%) and residual set (392 admissions/patients, 15%).

For the 392 patients of the residual set, physicians' notes were manually inspected to prevent the inclusion of falsely documented, implausible cases in the test set, which led to the exclusion of 28 patients. From the revised data set, 350 patients were randomly chosen and composed the final test set for evaluation of the algorithm.

The training and test sets were highly similar with respect to the baseline characteristics and the patients' total observation times, which are presented in Table 2 and Fig. 2 (Rank et al., 2020).

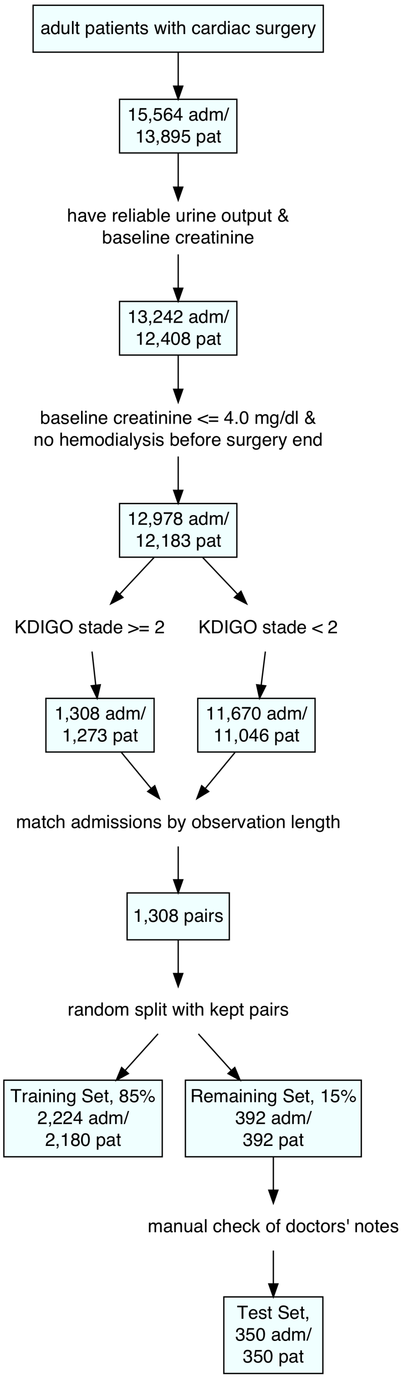

adult patients with cardiac surgery
15,564 adm/ 13,895 pat
have reliable urine output & baseline creatinine
13,242 adm/ 12,408 pat
baseline creatinine <= 4.0 mg/dl & no hemodialysis before surgery end
12,978 adm/ 12,183 pat
KDIGO stade >= 2
KDIGO stade < 2
1,308 adm/ 1,273 pat
11,670 adm/ 11,046 pat
match admissions by observation length
1,308 pairs
random split with kept pairs
Training Set, 85% 2,224 adm/ 2,180 pat
Remaining Set, 15% 392 adm/ 392 pat
manual check of doctors' notes
Test Set, 350 adm/ 350 pat

Figure 1. Patient selection process. adm = admissions, pat = patients. Reprinted from Fig. 3 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

Table 2. Baseline characteristics across the training and the test set. AKI = acute kidney injury, CPB = cardiopulmonary bypass. Reprinted from Supplementary Table 1 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

	Training Set	**Test Set**
No. of admissions	2224	350
No. of individual patients (%)	2180 (98)	350 (100)
No. of cases with AKI (%)	1112 (50)	175 (50)
Length of observation period in days, median (interquartile range)	1.33 (0.58 - 2.36)	1.30 (0.55 - 2.13)
Age, median (interquartile range)	72 (60 - 79)	71 (61 -79)
Male, No. (%)	1424 (64)	233 (67)
Baseline creatinine [mg/dl], median (interquartile range)	1.1 (0.83 - 1.4)	1.0 (1.0 - 1.0)
Baseline urea [mg/dl], median (interquartile range)	43 (32 - 62)	41 (31 - 56)
Time in operation theatre [minutes] median (interquartile range)	308 (208 - 450)	314 (214 - 428)
On-pump procedures, No. (%)	1134 (51)	177 (51)
Aortic cross clamp time [minutes], median (interquartile range)	81 (53 - 105)	77 (52 - 100)
CPB time [minutes], median (interquartile range)	118 (78 - 183)	122 (83 - 180)

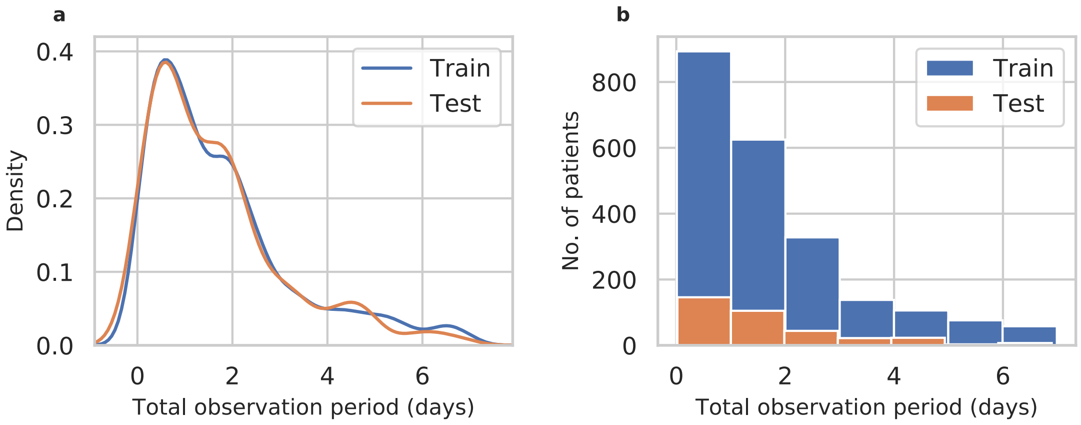


Figure 2. Length of observation windows of the patients in the training and in the test set. (a) density distribution. (b) histogram. Reprinted from Fig. 4 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

2.3 The basic principle of machine learning

Machine learning as an application of AI denotes the automated generation of knowledge from experience. An ML system autonomously learns patterns from examples without being explicitly instructed by humans on how these patterns look like. Multiple types of ML exist (*e.g., supervised*, *unsupervised* and *reinforcement learning*, see Fig. 3). In the study carried out for this doctoral project, *supervised learning* was applied. For a given set of input parameters (features), the algorithm produces an output (label). In supervised learning, the desired output is known during the training phase. Thus, the output of the algorithm can be compared with the correct output, i.e. the learning is *"supervised"*.

In the study, the outcome of interest was the development of AKI KDIGO stage 2 or 3 within seven days after cardiothoracic surgery. The set of input parameters comprised the time series of 96 routinely measured variables that were recorded in the EHR system and are further described in the following section.

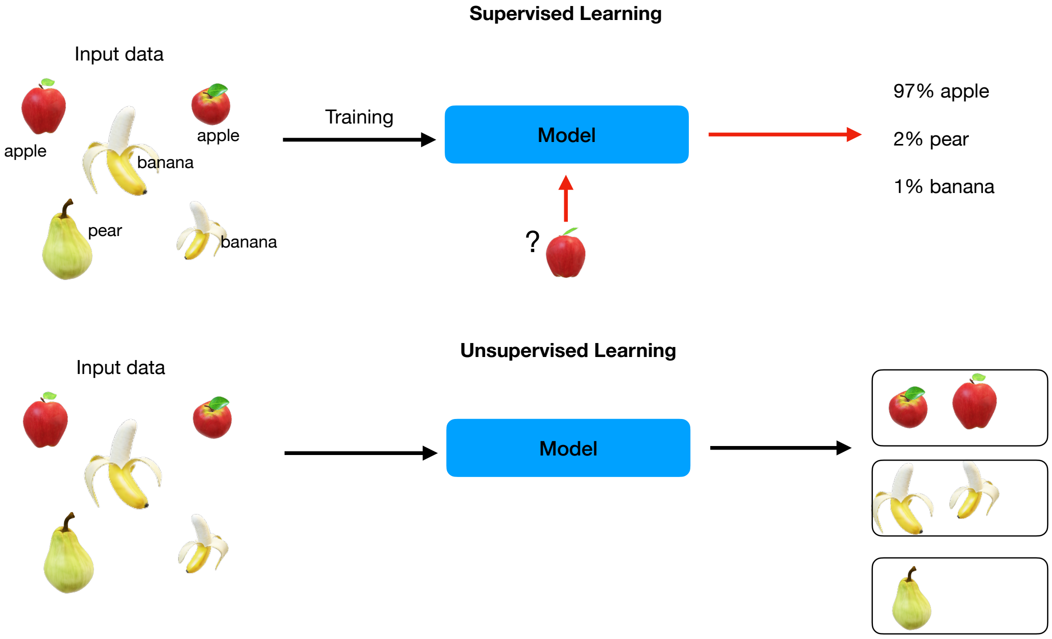


Figure 3. Supervised vs unsupervised machine learning. In supervised learning, the input data is labelled (annotated) for the training process. The model learns the specific characteristics (*e.g.*, shape, color) of the different labels (apple/banana/pear). For new data, the model predicts these labels. In unsupervised learning, labels are not known in advance. The algorithm tries to classify the data into groups with common characteristics (*e.g.*, shape, color).

2.4 Feature selection and data preprocessing

As input parameters for the model, 96 routinely collected parameters from the EHR system were selected and are shown in Table 3 (Rank et al., 2020). Most of the input features were of dynamic nature and, thus, can change over time. The last creatinine/urea value before the operation - or, in the case of absence, the first postoperative value - was defined as baseline creatinine/urea.

One very sensible parameter is the urine flow. First, it defines one criterion of AKI (< 0.5ml/kg/h for ≥ 12 hours for stage 2, see Table 1). However, on normal wards, the documentation of the urine flow was observed to be often insufficient - potentially with autonomous and mobile patients that did not report their urine output to clinical staff. To avoid the risk of false-positive AKI labels, the AKI urine criterion was only included in the AKI label definition while a patient was treated in an ICU/recovery room but not on normal wards.

Moreover, 22 frequently administered agents were incorporated, which were reported to have nephrotoxic effects (Kitano et al., 2014; Koch et al., 2008; Mazer & Perrone, 2008; Naughton, 2008; Nuis et al., 2012; Redondo-Pachon et al., 2014).

Table 3. Input feature overview. Adapted from Table 5 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Adapted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

Feature Group (no. features)	**Features**
Patient characteristics (4)	Age, sex, weight, height
Laboratory results (25)	Phosphate, total bilirubin, baseline creatinine, creatinine, baseline urea, urea, GFR, creatine kinase (CK), CK-MB, red blood count, white blood count, platelets, C-reactive protein (CRP), gamma-glutamyltransferase, glutamic oxaloacetic transaminase, hemoglobin, international normalized ratio, lactate dehydrogenase (LDH), magnesium, hematocrit, prothrombin time, partial thromboplastin time, mean corpuscular hemoglobin, mean corpuscular volume, mean corpuscular hemoglobin concentration
Surgery procedure (17)	Valve surgery, transcatheter aortic valve implantation (TAVI), endovascular TAVI, transapical TAVI, coronary artery bypass grafting (CABG), off-pump CABG, aortic surgery, assist device, ventricular assist device, extracorporeal membrane oxygenation system (ECMO), endovascular aortic stent implantation, transplantation, other major major cardiac surgery, isolated other major cardiac surgery, transcatheter mitral valve implantation (TMVI), endovascular TMVI, transapical TMVI (from logistic regression text model)
Further surgery characteristics (3)	Aortic cross-clamp time, cardiopulmonary bypass time, time in operation theatre
Vital signs (8)	Systolic, mean and diastolic arterial pressure, central venous pressure (CVP), heart frequency, pulse, body temperature, oxygen saturation
Arterial blood gas values (BGA) (15)	Base excess, bicarbonate, glucose, hemoglobin, oxygen saturation, partial pressure of carbon dioxide and oxygen, total carbon dioxide, pH level, potassium, sodium, calcium, lactate, carboxyhemoglobin, oxyhemoglobin
Fluid output (2)	Bleeding rate, urine flow rate
Nephrotoxic agents (22)	Allopurinol, Aminoglycosides, Amphotericin B, Antiplatelet agents (Clopidogrel, Ticlopidine), Benzodiazepines, Cephalosporins, Cyclosporine, Haloperidol, Ketamine, Nonsteroidal anti-inflammatory drugs, Paracetamol, Penicillines, Proton pump inhibitors, Pyrazolone derivatives, Quinolones, Ranitidine, Rifampin, Sulfonamides, Tacrolimus, Val/ganciclovir, Aciclovir, Vancomycin, red blood cell transfusions

One variable that was assumed to have a major impact on the AKI prediction, but could not be used in its raw form, was the type of surgery. It was documented in the EHR partly in free text form and partly in predefined categories. The algorithm chosen for AKI prediction was not designed in a way to interpret unstructured text in the first place but demanded either categorical or continuous input features. Therefore, a separate set of 17 logistic regression models was developed beforehand. It took both types of text information (categorical and free text) as input and derived for 17 predefined operation types a probability that a patient underwent the respective surgery. These probabilities were then used as continuous features for the final prediction model.

The AKI prediction model developed in the study received 15-minute intervals of all selected features as input. Forward imputation was performed to fill missing values (exception: nephrotoxic agents). In case of the absence of a precedent value, pre-selected default values were imputed.

The exact effect and duration of action of a drug are difficult to assess as both depend on the used excipients, drug-drug-interactions, dosage and a patients' metabolism. For this reason, medication treatment was encoded as follows: Whenever a nephrotoxic drug was administered, the value for the respective drug feature was set to 1 merely at the time slice that followed the administration. The values for this drug at the other time slices were set to 0.

To improve the speed of convergence of the algorithm, all features but the surgery types were scaled as follows (LeCun et al., 2012):

$$X_{scaled} = \frac{X-\mu(X_{train})}{IQR(X_{train})} \tag{1}$$

with $\mu(X_{train})$ representing the median and $IQR(X_{train})$ the interquartile range of the feature X in the training set.

The patient selection process, the feature preprocessing, and the data imputation was conducted with R v3.3.3 (R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/) and Python v3.6.7 (The Python Software Foundation, Beaverton, OR) with packages *IPython* (Perez & Granger, 2007) (v7.5.0), *Numpy* (van der Walt et al., 2011) (v1.16.2), *Pandas* (McKinney & Others, 2010) (v0.24.2), *Scikit-learn* (Pedregosa et al., 2011) (v0.19.1) and *Matplotlib* (Hunter, 2007) (v3.1.0).

2.5 Modelling

There are a variety of different ML models available (*e.g.*, Support Vector Machines, Logistic Regression, AdaBoost, Decision Trees). These models are suitable tools for static prediction but do not intrinsically assess the temporal evolution of parameters. Models with this ability are recurrent neural networks (RNN) which embed information about preceding time points and connect single timesteps (see Fig. 4 (Rank et al., 2020)).

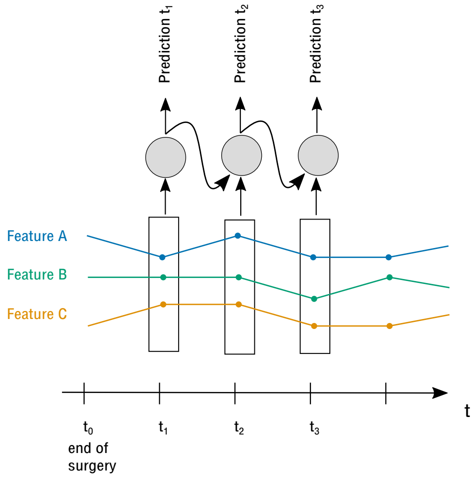


Figure 4. The basic principle of a recurrent neural network (RNN). The input to the RNN at each time step comprises the features of the respective time step, as well as the output obtained from the preceding time slice. Reprinted from Fig. 5 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

Thus, a set of RNNs with various building blocks (*e.g.*, a preceding convolutional layer, various cell types) was developed that predicted the risk of AKI every 15 minutes during a patient's observation window.

The hyperparameters were tuned by 5-fold stratified cross-validation on only the training set using the Adam optimizer (Kingma & Ba, 2014) and a static learning rate of 0.001. As the target parameter, the highest overall AUC was chosen. The final models were then evaluated on the unseen test set.

The hyperparameters of an RNN are influenced by the initialization and the sequence of presented training samples. For this reason, ten models were developed with identical hyperparameters but varying initializations. A final prediction was then obtained by combining the predictions of the ten models.

Different RNN architectures were tested, but only small differences in the AUC could be observed. Overall, gated recurrent unit cells (GRU) (Cho et al., 2014) tended to outperform long-short-term memory (LSTM) cells (Zaremba et al., 2014). A convolutional layer before the RNN did not increase the AUC. As the time sequences were relatively long (max. 673 time slices), also the phased LSTM cell was implemented, which is supposed to enhance the performance of a model that has to learn from very long sequences (Neil et al., 2016). It led to a lower AUC. Therefore, a model consisting of one layer of 100 neurons, the GRU cell and an output layer with softmax function was chosen as the final RNN.

The modelling part was performed using Python v3.6.7 (The Python Software Foundation, Beaverton, OR) with packages *IPython* (Perez & Granger, 2007) (v7.5.0), *Pandas* (McKinney & Others, 2010) (v0.24.2) and *Numpy* (van der Walt et al., 2011) (v1.16.2), *Scikit-learn* (Pedregosa et al., 2011) (v0.19.1), *Matplotlib* (Hunter, 2007) (v3.1.0) and *Tensorflow* (Abadi et al., 2016).

2.6 Evaluation of the RNN performance

2.6.1 Statistical measures

The following statistical measures were calculated to assess the performance of the RNN based on the independent test set:

- area under the curve (AUC)
- precision-recall-AUC (PR_AUC)
- accuracy
- sensitivity
- specificity
- positive predictive value (PPV)
- negative predictive value (NPV)
- false-positive rate (FPR)
- F_1-score
- mean of the Brier score (Brier, 1950) $\overline{MSE}_{pat}$

The mean squared error MSE_{pat} , or Brier score, of a single patient *j* is determined as follows:

$$MSE_{pat} = 1/ts_j \sum_{i=0}^{ts_j} (y_{ji} - y_{jt})^2 \quad (2)$$

where ts_j denotes the number of timesteps, y_{ji} the predicted value at time point *i*, and y_{jt} the true class of patient *j* (AKI/non-AKI).

Thus, $0 \leq \overline{MSE}_{pat} \leq 1$, whereas 0 denotes perfect prediction and 1 the opposite classification. $\overline{MSE}_{pat}$ is the only of the determined statistics that is not influenced by the length of a patient's observation window and the number of available timesteps for the respective patient.

Accuracy, sensitivity, specificity, NPV, PPV, FPR and F_1-score require a threshold that divides the continuous prediction into positive and negative classifications. This threshold is rather arbitrary and was set to the value that resulted in a sensitivity of 0.85 in the training set.

2.6.2 Adjustment of confidence intervals

Multiple predictions over time for one individual patient are in general highly correlated. These predictions can be considered clustered predictions. Therefore, an adjustment of the confidence intervals of the determined statistical measures was required. The 95% confidence interval of statistical measure *X* was determined as follows:

$$X \; +- 1.96 \, \sigma(X)$$

Where $\sigma(X)$ denotes the standard error of variable *X* and

$$\sigma(X) \; = \; \sqrt{\frac{X(1-X)}{n_{eff}}} \tag{3}$$

The effective sample size n_{eff} was determined by accounting for intracluster correlation as follows (Kalton et al.):

$$n_{eff} \; = \frac{n}{DE} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m_i} 1}{DE} \tag{4}$$

k denotes the number of patients, m_i the number of time steps of patient *i*, and *DE* the design effect or variance inflation factor (Kerry & Bland, 2001):

$$DE \; = \frac{\overline{m} \, k}{\sum_{i=1}^{k} \frac{m_i}{1+(m_i-1)ICC}} \tag{5}$$

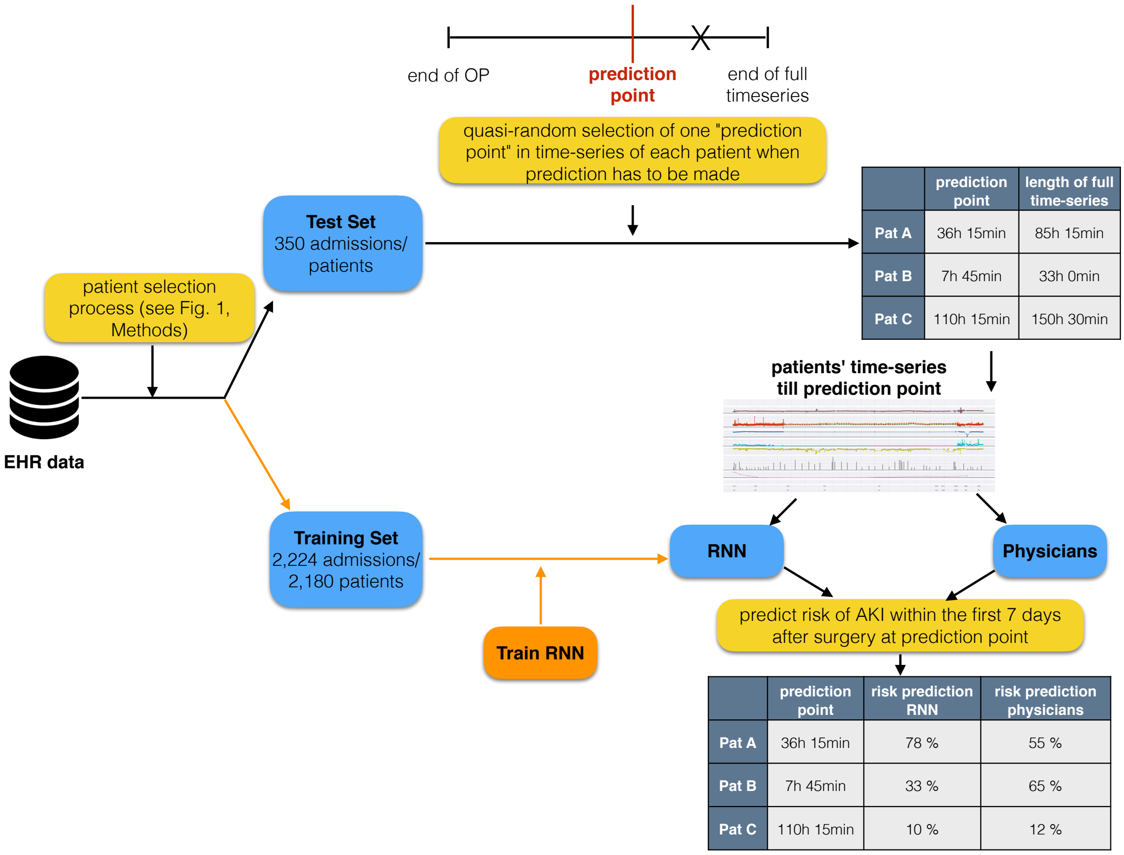
ICC refers to the intracluster correlation coefficient, calculated by the R package *ICC* (Wolak et al., 2012) (v2.3.0).

2.7 Comparing the RNN vs human performance

2.7.1 Experimental design

The second aim of the study underlying this doctoral project was to compare the performance of the RNN with that of experienced clinicians. The experimental design of this comparison is shown in Fig. 5 (Rank et al., 2020). For each clinical case in the test set, a *'prediction point'* in the patients' observation window was selected quasi-randomly. Quasi-random sampling is a method aiming at preventing cluster formations that can occur in real uniform random sampling

(Press et al., 1992; Weyl, 1916). Using this technique avoided the selection of prediction points lying in, *e.g.*, only the second half of patients' observation windows.


end of OP
prediction point
end of full timeseries
quasi-random selection of one "prediction point" in time-series of each patient when prediction has to be made
Test Set
350 admissions/ patients
patient selection process (see Fig. 1, Methods)
EHR data
patients' time-series till prediction point
Training Set
2,224 admissions/ 2,180 patients
Train RNN
RNN
Physicians
predict risk of AKI within the first 7 days after surgery at prediction point

	prediction point	length of full time-series
Pat A	36h 15min	85h 15min
Pat B	7h 45min	33h 0min
Pat C	110h 15min	150h 30min

	prediction point	risk prediction RNN	risk prediction physicians
Pat A	36h 15min	78 %	55 %
Pat B	7h 45min	33 %	65 %
Pat C	110h 15min	10 %	12 %

Figure 5. Study design for the head-to-head comparison recurrent neural network (RNN) vs physicians. A training set and a test set were compiled from the electronic health record (EHR) data. The RNN was trained on the training set (orange path). The test set was used for evaluation. In each of its patient's observation period, a *'prediction point'* was selected quasi-randomly. Physicians and the RNN received the EHR data up to this *prediction point*, whereas all information collected after the *prediction point* (marked as **X**) was hidden. Both had to forecast postoperative AKI at the *prediction point*. Adapted from Fig. 1 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Adapted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

At this *prediction point*, a prediction of whether the respective patient would develop AKI within the first seven postoperative days had to be made - both by the RNN and by a physician.

The physicians received all time series data up to this *prediction point* presented on a screen, similar to an electronic patient chart. In contrast to the RNN, the time series information was displayed in the originally detected time resolution (up to 1 minute) and the physicians were given information about all administered drugs - not only about the nephrotoxic agents. Furthermore, they received the surgery type as unstructured text that was manually derived from physicians' notes, and they were informed about the 50% incidence rate of AKI in the test set.

For each patients' *prediction point*, both the RNN and a physician made a risk prediction *r* (ranging from 0% to 100%). The physicians additionally had to give a binary classification (AKI/non-AKI) for the respective case.

In total, seven physicians of different experience levels (from senior resident to senior consultant) participated in the study, each making predictions for 50 different patients. All participating physicians had at least one year of working experience on a cardiothoracic ICU and a total of at least five years of clinical experience.

2.7.2 Sample size calculation and statistical comparison

The head-to-head comparison was designed as a non-inferiority experiment. The goal was to prove that the RNN could predict AKI not significantly worse than experienced clinicians.
A predictive quality score *S* for a single prediction was defined based on the predicted probability *r* as follows:

$S = r$, in case the patient developed AKI (6)

$S = 1 - r$, in case the patient did not develop AKI (7)

S was non-normally distributed for the RNN's predictions and was then transformed for power analysis and sample size calculation as follows:

$$X = -log(-log(S)) \tag{8}$$

to reach an approximately normal distribution of *X*. It was assumed that *X* would also follow a normal distribution for the physicians' predictions.

A significance level of α = 0.025, a power ≥ 80% and a non-inferiority margin δ = 0.3 lead to a sample size of n = 350 patients. The non-inferiority margin δ = 0.3 was equivalent to allowing sensitivity+specificity of the RNN to be a maximum of 5.5% smaller than of the physicians' predictions.

AUC, PR_AUC, accuracy, sensitivity, specificity, PPV, NPV, FPR, Brier score and F_1-score were then determined for the predictions of the RNN and the physicians. Here, the threshold dividing between positive and negative class was set to 0.5 as this was the naturally occurring threshold of the physicians' predictions that reflected the binary 'AKI/non-AKI'-prediction. The calculation of confidence intervals was identical to that described in Chapter 2.6.2. As only one prediction was made for each patient: n_{eff} = n = 350.

For the comparison of the RNN's and the physicians' performance, a significance level of α = 0.05 was set. The predictive quality score *S* was compared between the RNN and the physicians by the paired t-test. The comparison of the two receiver operating characteristics (ROC) curves was conducted using DeLong's (DeLong et al., 1988) method for correlated receiver operating characteristics (ROC) and the R-package *pROC* (v1.9.1) (Robin et al., 2011). The calibration of both predictors (RNN and physicians) was analysed with the Hosmer-Lemeshow-Test (Jr. et al., 2013) and the R package *ResourceSelection* (v0.3-2) (Lele et al., 2016).

3. Results

3.1 Predictive performance of the RNN

The performance of the final RNN assessed on the independent test set of n = 350 patients is presented in Table 4 (Rank et al., 2020). The RNN reached an AUC (confidence interval (CI)) of 0.893 (0.862 - 0.924), a PR_AUC of 0.903 (0.873 - 0.933) and an accuracy of 0.825 (0.786 - 0.863) based on a threshold of 0.41 that led to a sensitivity of 0.85 in the training set.

The model performance on an imbalanced test set (AKI incidence rate of 10%) is shown in Table 5 (Rank et al., 2020). The observed AUC was around five percentage points lower than for the test set with balanced class proportions. The training of the model was performed with a balanced data set. Thus, as expected, the FPR went up when testing the RNN on a set with only a 10% incidence of AKI. Correspondingly, the NPV surpassed 99%.

Table 4. Model performance metrics for the balanced test set (n = 350 admissions/patients). Acc = accuracy, AUC = area under the curve, CI = confidence interval, F_1 = F_1-score, FPR = false-positive rate, $\overline{MSE}_{pat}$ = mean of the brier score of each patient, NPV = negative predictive value, PPV = positive predictive value, PR_AUC = precision-recall AUC, Sens = sensitivity, Spec = specificity. Reprinted from Table 1 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

Threshold-independent metrics (95 % CI)			Metrics at a fixed sensitivity of 0.85 (95 % CI)						
AUC	PR_AUC	$\overline{MSE}_{pat}$	Acc	Sens	Spec	F_1	FPR	NPV	PPV
0.893 (0.862 - 0.924)	0.903 (0.873 - 0.933)	0.124 (0.090 - 0.159)	0.825 (0.786 - 0.863)	0.853 (0.802 - 0.904)	0.798 (0.741 - 0.855)	0.826 (0.776 - 0.876)	0.202 (0.145 - 0.259)	0.851 (0.799 - 0.903)	0.801 (0.745 - 0.857)

Table 5. Model performance metrics of an imbalanced test set (n = 1945 admissions/patients). The incidence rate of 10% acute kidney injury in this test set with n = 1945 admissions corresponds to that of the original study population. Metrics that depend on a threshold that discriminates between positive and negative classes are presented at a fixed sensitivity of 0.85. Acc = accuracy, AUC = area under the curve, CI = confidence interval, F_1 = F_1-score, FPR = false-positive rate, $\overline{MSE}_{pat}$ = mean of the Brier score of each patient, NPV = negative predictive value, PPV = positive predictive value, PR_AUC = precision-recall AUC, Sens = sensitivity, Spec = specificity. Reprinted from Table 1 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

Threshold-independent metrics (95 % CI)			Metrics at a fixed sensitivity of 0.85 (95 % CI)						
AUC	PR_AUC	$\overline{MSE}_{pat}$	Acc	Sens	Spec	F_1	FPR	NPV	PPV
0.846 (0.831 - 0.862)	0.152 (0.137 - 0.168)	0.153 (0.137 - 0.169)	0.747 (0.728 - 0.765)	0.850 (0.768 - 0.932)	0.743 (0.724 - 0.762)	0.191 (0.159 - 0.222)	0.257 (0.238 - 0.276)	0.993 (0.988 - 0.997)	0.107 (0.082 - 0.132)

3.2 Comparing the RNN to human prediction

3.2.1 Overall performance

The comparison of the performance of the physicians and the RNN is displayed in Table 6 (Rank et al., 2020). (The values of the RNN in Table 6 differ slightly from those in Table 4, as in this experiment, only one *prediction point* was tested for each patient. In contrast, in the full RNN evaluation in Chapter 3.1, all predictions for all time points of the observation window of all patients were included.)

Table 6. Performance metrics of the recurrent neural network (RNN) and the physicians on the balanced test set (n = 350 admissions/patients). Acc = accuracy, AUC = area under the curve, Brier = Brier score, CI = confidence interval, F_1 = F_1-score, FPR = false-positive rate, NPV = negative predictive value, PPV = positive predictive value, PR_AUC = precision-recall AUC, Sens = sensitivity, Spec = specificity. Reprinted from Table 2 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

	Threshold-independent metrics (95 % CI)			**Metrics based on a threshold of 0.5 for positive/negative classification (95 % CI)**						
	AUC	**PR_AUC**	**Brier**	**Acc**	**Sens**	**Spec**	**F_1**	**FPR**	**NPV**	**PPV**
RNN	0.901 (0.870 - 0.932)	0.907 (0.877 - 0.937)	0.122 (0.088 - 0.156)	0.846 (0.808 - 0.884)	0.851 (0.798 - 0.904)	0.840 (0.787 - 0.894)	0.847 (0.797 - 0.897)	0.160 (0.106 - 0.214)	0.850 (0.797 - 0.903)	0.842 (0.788 - 0.896)
Physicians	0.745 (0.699 - 0.791)	0.747 (0.701 - 0.793)	0.217 (0.174 - 0.260)	0.711 (0.664 - 0.759)	0.594 (0.521 - 0.667)	0.829 (0.773 - 0.884)	0.673 (0.609 - 0.738)	0.171 (0.116 - 0.227)	0.671 (0.601 - 0.741)	0.776 (0.715 - 0.838)

The RNN surpassed the physicians across all performance metrics. It yielded an AUC of 0.901 (0.870 - 0.932), whereas the physicians only reached an AUC of 0.745 (0.699 - 0.791). DeLong's test for correlated ROC curves showed a significant superiority of the RNN ($p <$ 0.001, Z = 6.85)). Additionally, the paired t-test revealed a significantly higher mean of the predictive quality score *S* for the RNN (RNN: 0.754 vs physicians: 0.639 (0.754 vs 0.639, *p* < 0.001, *t*-statistic = 8.47, df = 349). Fig. 6a and 6b show the ROC curves and the precision-recall curves.

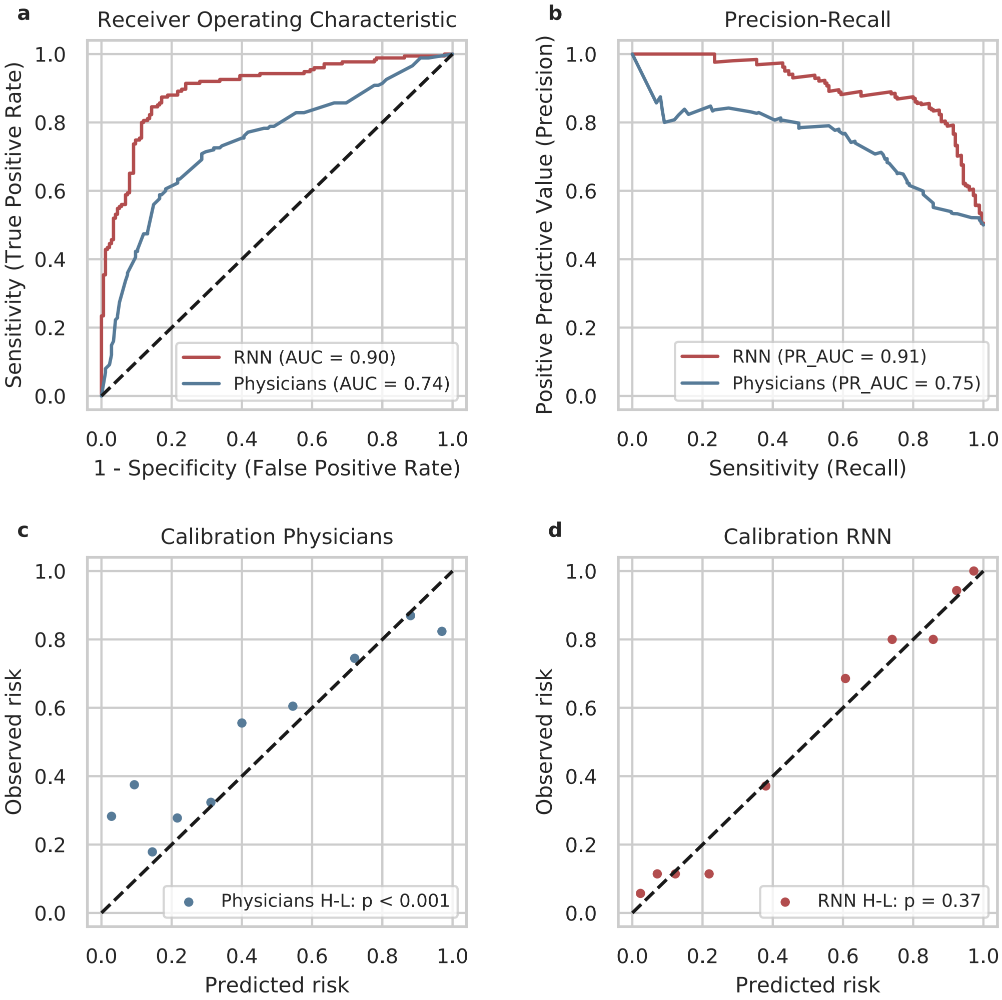


Figure 6. Discrimination and calibration of the recurrent neural network (RNN) and the physicians. (a) receiver operating characteristics (ROC), (b) precision-recall curve, (c) calibration of the physician's predictions, (d) calibration of the RNN's predictions. AUC = area under the curve. H-L = Hosmer-Lemeshow-Test, PR_AUC = precision-recall AUC. Reprinted from Fig. 2 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

Moreover, an evaluation of the calibration of the physicians' and the RNN's predictions was performed. Calibration characterizes the agreement between the frequencies of the observed events and the predictions. In a calibration plot, perfect calibration would show as two points - one at (0,0) and the other at (1,1), as a perfect model would always forecast 0 for negative and 1 for positive outcomes. In an imperfectly but well-calibrated model, all points should be located on the diagonal between (0,0) and (1,1). Then the observed frequencies correspond to the predicted frequencies of events. Fig. 6c displays the calibration of the physicians' predictions. In the graph sections with high predicted risks, the predicted frequencies widely agree with the observed event frequencies. However, for various AKI-cases, physicians predicted lower AKI risks, resulting in false-negative predictions (lower left part of the graph). The Hosmer-Lemeshow test indicated that the physicians' risk assessment was not well calibrated (H_0: The predictions fit the observed data well, H_1: The predictions do not fit the observed data well, $p < 0.001$, $X^2 = 165.5$, df = 8). In comparison, Fig. 6d illustrates that the RNN's predictions were well calibrated as all points are located on or close to the diagonal ($p = 0.37$, $X^2 = 8.67$, df = 8). This is also the case for the intervals of low predicted risks.

3.2.2 Time-dependent performance

Additionally, the RNN's and physicians' predictive performance was evaluated at various time points preceding the outcome of a patient (AKI vs non-AKI/discharge). The results are shown in Table 7 (Rank et al., 2020). In the case of a long period of time between the prediction point and the event, both RNN and physicians generally predicted less accurately. However, they also tended to perform worse if the event was very close in time (≤ 2h). As the median observation time in this group of patients was also quite short, it can be assumed that the respective AKI-patients in this group developed the AKI quickly after the operation. Most likely, neither the RNN nor the physicians received sufficient information to reliably forecast AKI in that situation. Still, even with this group, the RNN yielded a sensitivity of 0.789 (vs 0.632 for the physicians).

Table 7. Performance metrics of recurrent neural network (RNN) and physicians in temporal dependence to the event. Acc = accuracy, AKI = number of patients with acute kidney injury, AUC = area under the curve, Brier = Brier score, F_1 = F_1-score, FPR = false-positive rate, MOL = median total observation length, NPV = negative predictive value, PPV = positive predictive value, PR_AUC = precision-recall AUC, Sens = sensitivity, Spec = specificity. Reprinted from Table 3 from *"Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance"* by N. Rank et al., 2020, NPJ Digital Medicine, *3*, 139. Reprinted with permission under a Creative Commons Attribution 4.0 International License, accessed http://creativecommons.org/licenses/by/4.0/.

	Time to event	patients	AKI	MOL	AUC	PR_ AUC	Brier	Acc	Sens	Spec	F_1	FPR	NPV	PPV
RNN	0h to 2h	54	19	8.3h	0.913	0.837	0.113	0.870	0.789	0.914	0.811	0.086	0.889	0.833
Physicians	0h to 2h	54	19	8.3h	0.709	0.552	0.199	0.759	0.632	0.829	0.649	0.171	0.806	0.667
RNN	2h to 6h	63	29	12.5h	0.881	0.88	0.13	0.825	0.862	0.794	0.820	0.206	0.871	0.781
Physicians	2h to 6h	63	29	12.5h	0.853	0.861	0.152	0.794	0.793	0.794	0.780	0.206	0.818	0.767
RNN	6h to 12h	63	34	17.8h	0.942	0.948	0.088	0.921	0.971	0.862	0.930	0.138	0.962	0.892
Physicians	6h to 12h	63	34	17.8h	0.811	0.798	0.19	0.746	0.618	0.897	0.724	0.103	0.667	0.875
RNN	12h to 24h	74	42	36.4h	0.888	0.921	0.128	0.824	0.881	0.750	0.851	0.250	0.828	0.822
Physicians	12h to 24h	74	42	36.4h	0.693	0.706	0.257	0.689	0.667	0.719	0.709	0.281	0.622	0.757
RNN	24h to 48h	60	31	46.4h	0.890	0.899	0.142	0.817	0.774	0.862	0.814	0.138	0.781	0.857
Physicians	24h to 48h	60	31	46.4h	0.718	0.774	0.246	0.633	0.387	0.897	0.522	0.103	0.578	0.800
RNN	48h to 168h	36	20	99.0h	0.875	0.929	0.132	0.806	0.750	0.875	0.811	0.125	0.737	0.882
Physicians	48h to 168h	36	20	99.0h	0.647	0.741	0.274	0.611	0.400	0.875	0.533	0.125	0.538	0.800

4. Clinical applications, prospective research questions, limitations and ethical considerations

4.1 Clinical applications and prospective research questions

The model developed in this doctoral project could serve as a real-time monitoring tool that recognizes the risk of kidney failure and could, thus, help to improve patient care by informing physicians at an early stage of imminent AKI. Due to increasing digitalization, it is conceivable that soon every physician will be equipped with a personal mobile device (*e.g.*, smartwatch, tablet). An overview chart could be displayed to each physician that shows the AKI risk for all their patients. Additionally, warning messages could be sent to the devices as soon as a certain customized risk threshold (based on a fixed value or a trend analysis) for a patient is surpassed. The notifications could further be extended by, *e.g.*, diagnostic checklists to facilitate systematic evaluation even in stressful situations. Furthermore, alerts could automatically be delivered to a nephrologist to allow further investigation of the case by a specialist. To think even further, a transfer of the model from a risk prediction tool to an application suggesting further diagnostic steps or treatment options is also imaginable. These clinical applications show tremendous potential, but they also pose a plethora of ethical and legal questions, which are discussed in Chapter 4.3.

The study carried out in the context of this dissertation is a retrospective study. All used information was drawn from the EHR of a single clinical center alone. Further studies are needed to confirm the performance of such an automated system on external data. Additionally, it should be investigated if a system similar to the one proposed here would be accepted and used by clinical staff. In prospective studies, it should be examined if the early prediction of AKI, and following interventive measures, can actually avert AKI and its consequences described.

Currently, the model is limited to the prediction of AKI. The ability to predict other endpoints (*e.g.*, pneumonia, sepsis) would represent a useful further development of the model.

The proposed model was developed on a training cohort of only patients admitted to a cardiothoracic surgery unit. It should further be investigated how such a model would perform on other patient cohorts (*e.g.*, patients with abdominal surgery, patients on a general ICU) as those patients probably have disparate risk profiles.

4.2 Limitations

The study carried out for this doctoral project has certain limitations. The prediction window was limited to the first seven postoperative days, whereas its length differed largely among the studied patients. The majority of patients had an observation time of fewer than three days, and only a small number of patients were observed up to seven days.
The implementation of the administered drugs in the developed model is highly simplistic as only the administration itself is assessed. No information about dosage, application length or route (p.o./i.v./…) was given to the model. These features could, however, be meaningful and ameliorate the prognostic capacity of the model.

In the retrospective study, physicians did not have the possibility to assess patients physically. Meaningful information like full internal status or further diagnostic results (*e.g.*, electrocardiogram, ultrasound) was neither given to the physicians nor to the RNN. Predicting a complication merely by parameters displayed on a screen does not reflect a physician's normal way of working, which could be a reason for the overall low predictive performance of the participating physicians.

Unlike simple linear models such as logistic regression models, RNNs are subject to a complex architecture. A high prediction accuracy often comes at the cost of the explainability of the model (Caruana et al., 2015). This makes it difficult to determine the exact causes of the imminent AKI. Patients who appear in good clinical condition are usually not automatically presumed to be at high risk of impending renal failure. If the model predicts high AKI risks for such patients, it would be essential to know the precise causes of the prediction in order to take the necessary preventive measures. At present, the model presented here serves as an early warning system. A more detailed evaluation of the causes of AKI should then be carried out by a specialist, i.e. a nephrologist.

From a developer's perspective, the integration of the proposed model into digital medical record systems is straightforward. The model only employs routinely recorded data, and all information is obtained and processed automatically. The actual difficulties of a real-world implementation, however, range from data privacy issues when patient data is transferred to third-party systems, technical limitations and business interests that may collide with one another.

4.3 Ethical considerations regarding the practical application of machine learning models in medicine

The application of AI to medical questions raises various ethical concerns that can be split into three major categories, namely, the acquisition of data, the model development and the clinical application of the respective models (Vayena et al., 2018).

4.3.1 Data acquisition

The data employed by ML models are covered by data protection regulations. In the European Union, the General Data Protection Regulation (GDPR) was adopted for this purpose (2018 Reform of EU Data Protection Rules, 2018). It states that informed consent of the concerned subjects is required in case their data is used and confers various rights on individuals, which have to be complied with by the parties who employ their data (McCall, 2018). Developers of ML models have to ensure that the required consents have been obtained. However, it is not always trivial to decide for what specific intent permission was granted (Vayena et al., 2018). Since ML usually requires very large amounts of data - often several thousand patient cases - it is hardly possible to elicit these aspects for each individual patient, let alone discuss them with each patient individually. Patients should therefore be made aware upon admission to the hospital that their collected data may also be used for ML applications and should have the opportunity to object to this.

4.3.2 Model development

Apart from data privacy issues, the development of ML models also raises ethical aspects that should be considered.

4.3.2.1 Target population and outliers

ML algorithms are usually trained and evaluated on larger sets of patients. Thus, they are highly influenced by the “broad mass” of the patients in the training set. First of all, it should be ensured that the target population is similar to the training population (Vayena et al., 2018). For instance, a model that was trained on mostly middle-aged men might not be suitable for making risk predictions for senior women. With the increasing amount of digitalization, we can, though, expect that most large patient groups will be covered with enough training samples in the near future.

More difficult is the handling of “outliers”. Outlier patients in this context could be ethnic minorities or patients with rare diseases that might have highly different risk profiles than the patients in the training population. Thus, end-users (clinical staff) should be aware of the population on which the respective model was trained and for which cases its predictions must be particularly critically scrutinized.

4.3.2.2 Legal liability

The next question that arises is who is liable in case the model fails (Vayena et al., 2018). The model developers are not at the end of the decision chain, and the ultimate responsibility for medical decisions rests with the physician. However, the individual physician was not usually involved in the model development process. Therefore, it seems difficult to hold the physician liable in the case of technical errors in the model. In individual cases, it may be difficult to decide to what extent it was reasonable to rely on the prediction of an ML or at what point a physician should have suspected a technical error in the model. However, this problem also applies to other diagnostic test procedures and is not specific to ML algorithms.

4.3.3 Clinical deployment of the model

As soon as a machine learning model is firmly integrated into everyday clinical practice and is not only used for research purposes, further ethical aspects should be considered.

4.3.3.1 Responsible decision making and patients’ autonomy

First of all, it should be ensured that clinical staff have at least a fundamental understanding of the model logic and the pitfalls of the model so that they can make responsible clinical decisions. If this is not the case, and physicians are not able to explain their decisions to their patients, the relationship of trust between physician and patient could be disturbed. Moreover, the autonomy of the patient could be violated, as an informed decision presupposes that the respective patient has sufficient information about his/her state of health. Simply informing a patient about the predictive value of the ML model without its origin may not be sufficient for some patients (Vayena et al., 2018).

4.3.3.2 Critical evaluation of model predictions

Introducing a prediction tool with generally high accuracy always carries the risk that end users will rely too much on its predictions and lose their common sense and instinct. This may be less the case for physicians who have already gained many years of clinical experience without such models and may therefore be more critical of a model's predictions. However, if inexperienced physicians rely too much on computer models, this could cause them to develop less ability to make independent clinical assessments of patients. This is made particularly difficult by the fact that a physician potentially has to justify every decision "against the model". Particularly with regard to a possible subsequent legal reappraisal of individual patient cases, this may lead to decisions against the model being avoided. The objective decision-making of a physician can, thus, be restricted by a prediction model.

4.3.3.3 False conclusions of true patterns

ML models learn patterns in the data without reviewing them for reasonableness which may result in making false conclusions of true patterns (Caruana et al., 2015). It is, for instance, possible that more intensive preventive measures are taken a priori for patients with known chronic kidney disease (CKD) than for patients without renal damage since CKD is a commonly known risk factor for AKI. These preventive measures (*e.g.*, a priori consultation of a nephrologist after surgery, avoidance of nephrotoxic medication) could lead to the phenomenon that patients with CKD develop AKI less often than patients without CKD. An ML algorithm might then learn that the risk for developing AKI is lower for patients with CKD than for patients without CKD, which clearly does not reflect the true underlying pathophysiology of AKI. End-users of the product should always be aware of the possibility of such erroneous conclusions, and in no case should they omit the enhanced a priori preemptive measures because of potential low-risk predictions for obvious high-risk patient groups. Otherwise, this could put high-risk groups at an even greater risk of developing complications.

4.3.3.4 Fair distribution of attention

Furthermore, if a machine learning model is used for monitoring and too much reliance is placed on it, there is a risk of "forgetting" patients for whom the model does not predict a high complication rate. This can lead to these patients being evaluated clinically less often by physicians. However, many prediction models are limited to predicting one or only a few complications. A low prediction probability for this one/few complications does, of course, not exclude that the respective patients develop other complications. If patients receive less attention due to low probabilities for the complication the model covers, their risk for other complications may increase. In addition, it should always be kept in mind that, as mentioned earlier, such models generally work well for the “broad mass” of patients but may fail for “outliers”.

4.3.3.5 Consideration of patients’ social circumstances

Most clinical prediction models are based primarily on physiological measures, diagnoses or other assessments of a patient's clinical condition. However, it is a physician’s role to also take into account patients’ social circumstances (*e.g.*, home care, compliance) when making decisions. The use of AI in medicine should not lead to less consideration of these components in physicians’ decisions; for instance, due to the fear of recourse claims by health insurers against hospitals if patients are hospitalized longer for social reasons contrary to the low-risk predictions made by an ML model.

To sum up, the points discussed in this section clearly show that ML models can be used as decision support for physicians but cannot and should not replace them.

5. Conclusions

In the context of this doctoral project, an RNN was constructed that predicted AKI within the first seven postoperative days with excellent accuracy. A head-to-head comparison with predictions by experienced physicians revealed that the RNN surpassed the latter with respect to all measured performance metrics. Integrating the model into existing digital medical record systems could help to forecast AKI before its onset and, thus, enable physicians to take preventive interventions at an early stage. The model employs routinely recorded medical data and therefore does not cause an extra burden for clinical staff. To leverage such models not only for research purposes but for real clinical use, further prospective studies are needed. In this regard, ethical aspects should be considered by both the model developers and the clinical staff.

6. References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwer, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. In *arXiv [cs.DC]*. arXiv. Last Retrieved October 23, 2021 from http://arxiv.org/abs/1603.04467v2

Aronson, S., Fontes, M. L., Miao, Y., Mangano, D. T., Investigators of the Multicenter Study of Perioperative Ischemia Research Group, & Ischemia Research and Education Foundation. (2007). Risk index for perioperative renal dysfunction/failure: critical dependence on pulse pressure hypertension. *Circulation*, *115*(6), 733–742.

Balasubramanian, G., Al-Aly, Z., Moiz, A., Rauchman, M., Zhang, Z., Gopalakrishnan, R., Balasubramanian, S., & El-Achkar, T. M. (2011). Early nephrologist involvement in hospital-acquired acute kidney injury: a pilot study. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, *57*(2), 228–234.

Ball, L., Costantino, F., & Pelosi, P. (2016). Postoperative complications of patients undergoing cardiac surgery. *Current Opinion in Critical Care*, *22*(4), 386–392.

Bennett, M., Dent, C. L., Ma, Q., Dastrala, S., Grenier, F., Workman, R., Syed, H., Ali, S., Barasch, J., & Devarajan, P. (2008). Urine NGAL predicts severity of acute kidney injury after cardiac surgery: a prospective study. *Clinical Journal of the American Society of Nephrology: CJASN*, *3*(3), 665–673.

Bihorac, A., Ozrazgat-Baslanti, T., Ebadi, A., Motaei, A., Madkour, M., Pardalos, P. M., Lipori, G., Hogan, W. R., Efron, P. A., Moore, F., Moldawer, L. L., Wang, D. Z., Hobson, C. E., Rashidi, P., Li, X., & Momcilovic, P. (2019). MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. *Annals of Surgery*, *269*(4), 652–662.

Brier, W. G. (1950). Verification of Forecasts Expressed in terms of probability. *Monthey Weather Review*, *78*(1), 1–3.

Burke-Gaffney, A., Svermova, T., Mumby, S., Finney, S. J., & Evans, T. W. (2014). Raised plasma Robo4 and cardiac surgery-associated acute kidney injury. *PloS One*, *9*(10), e111459.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730.

Cheng, P., Waitman, L. R., Hu, Y., & Liu, M. (2017). Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate? *AMIA Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2017*, 565–574.

Chertow, G. M., Lazarus, J. M., Christiansen, C. L., Cook, E. F., Hammermeister, K. E., Grover, F., & Daley, J. (1997). Preoperative renal risk stratification. *Circulation*, *95*(4), 878–884.

Chertow, G. M., Levy, E. M., Hammermeister, K. E., Grover, F., & Daley, J. (1998). Independent Association between Acute Renal Failure and Mortality following Cardiac Surgery. *The American Journal of Medicine*, *104*(4), 343–348.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *arXiv [cs.CL]*. arXiv. Last Retrieved October 23, 2021 from http://arxiv.org/abs/1406.1078v3

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55–63.

Costa e Silva, V. T., Liaño, F., Muriel, A., Díez, R., de Castro, I., & Yu, L. (2013). Nephrology referral and outcomes in critically ill acute kidney injury patients. *PloS One*, *8*(8), e70482.

Dąbrowska-Bender, M., Dykowska, G., Żuk, W., Milewska, M., & Staniszewska, A. (2018). The impact on quality of life of dialysis patients with renal insufficiency. *Patient Preference and Adherence*, *12*, 577–583.

Dasta, J. F., & Kane-Gill, S. (2019). Review of the Literature on the Costs Associated With Acute Kidney Injury. *Journal of Pharmacy Practice*, *32*(3), 292–302.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845.

Donchin, Y., & Seagull, F. J. (2002). The hostile environment of the intensive care unit. *Current Opinion in Critical Care*, *8*(4), 316-20.

2018 Reform of EU Data Protection Rules, European Commission, May 25, 2018.

Haase, M., Bellomo, R., Devarajan, P., Schlattmann, P., Haase-Fielitz, A., & NGAL Meta-analysis Investigator Group. (2009). Accuracy of neutrophil gelatinase-associated lipocalin (NGAL) in diagnosis and prognosis in acute kidney injury: a systematic review and meta-analysis. *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation*, *54*(6), 1012–1024.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How Many Variables Can Humans Process? *Psychological Science*, *16*(1), 70–76.

Hobson, C. E., Yavas, S., Segal, M. S., Schold, J. D., Tribble, C. G., Layon, A. J., & Bihorac, A. (2009). Acute kidney injury is associated with increased long-term mortality after cardiothoracic surgery. *Circulation*, *119*(18), 2444–2453.

Hobson, C., Ozrazgat-Baslanti, T., Kuxhausen, A., Thottakkara, P., Efron, P. A., Moore, F. A., Moldawer, L. L., Segal, M. S., & Bihorac, A. (2015). Cost and Mortality Associated With Postoperative Acute Kidney Injury. *Annals of Surgery*, *261*(6), 1207–1214.

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression: Third Edition* (pp. 157-169). In *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Huen, S. C., & Parikh, C. R. (2012). Predicting acute kidney injury after cardiac surgery: a systematic review. *The Annals of Thoracic Surgery*, *93*(1), 337–347.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Jayakumar, C., Ranganathan, P., Devarajan, P., Krawczeski, C. D., Looney, S., & Ramesh, G. (2013). Semaphorin 3A is a new early diagnostic biomarker of experimental and pediatric acute kidney injury. *PloS One*, *8*(3), e58446.

Kalton, G., Michael Brick, J., & Lê, T. (n.d.). *Chapter VI Estimating components of design effects for use in sample design*. Retrieved August 7, 2019, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.522.3221

Kerry, S. M., & Bland, J. M. (2001). Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in Medicine*, *20*(3), 377–390.

Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. (2012). KDIGO Clinical Practice Guideline for Acute Kidney Injury. *Kidney International. Supplement, 2,* 1–138.

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. Last Retrieved October 23, 2021 from http://arxiv.org/abs/1412.6980v9

Kitano, A., Motohashi, H., Takayama, A., Inui, K.-I., & Yano, Y. (2014). Valacyclovir-Induced Acute Kidney Injury in Japanese Patients Based on the PMDA Adverse Drug Reactions Reporting Database. *Drug Information Journal*, *49*(1), 81–85.

Koch, C. G., Li, L., Sessler, D. I., Figueroa, P., Hoeltge, G. A., Mihaljevic, T., & Blackstone, E. H. (2008). Duration of red-cell storage and complications after cardiac surgery. *The New England Journal of Medicine*, *358*(12), 1229–1239.

Koyner, J. L., Adhikari, R., & Edelson, D. P. (2016). Development of a multicenter ward–based AKI prediction model. *Clinical Journal of the American Society of Nephrology: CJASN*, *11*(11)*,* 1935–1943.

Koyner, J. L., Carey, K. A., Edelson, D. P., & Churpek, M. M. (2018). The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. *Critical Care Medicine*, *46*(7), 1070–1077.

Krawczeski, C. D., Goldstein, S. L., Woo, J. G., Wang, Y., Piyaphanee, N., Ma, Q., Bennett, M., & Devarajan, P. (2011). Temporal relationship and predictive value of urinary acute kidney injury biomarkers after pediatric cardiopulmonary bypass. *Journal of the American College of Cardiology*, *58*(22), 2301–2309.

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient BackProp. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade: Second Edition* (pp. 9–48). Springer Berlin Heidelberg.

Lele, S. R., Keim, J. L., & Solymos, P. (2016). ResourceSelection: resource selection (probability) functions for use-availability data. *R Package Version*, 2–6.

Mandelbaum, T., Scott, D. J., Lee, J., Mark, R. G., Malhotra, A., Waikar, S. S., Howell, M. D., & Talmor, D. (2011). Outcome of critically ill patients with acute kidney injury using the Acute Kidney Injury Network criteria. *Critical Care Medicine*, *39*(12), 2659–2664.

Mazer, M., & Perrone, J. (2008). Acetaminophen-induced nephrotoxicity: pathophysiology, clinical manifestations, and management. *Journal of Medical Toxicology: Official Journal of the American College of Medical Toxicology*, *4*(1), 2–6.

McCall, B. (2018). What does the GDPR mean for the medical community? *The Lancet*, *391*(10127), 1249–1250.

McIlroy, D. R., Wagener, G., & Lee, H. T. (2010). Neutrophil gelatinase-associated lipocalin and acute kidney injury after cardiac surgery: the effect of baseline renal function on diagnostic performance. *Clinical Journal of the American Society of Nephrology: CJASN*, *5*(2), 211–219.

McKinney, W., & Others. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Meersch, M., Schmidt, C., Hoffmeier, A., Van Aken, H., Wempe, C., Gerss, J., & Zarbock, A. (2017). Prevention of cardiac surgery-associated AKI by implementing the KDIGO guidelines in high risk patients identified by biomarkers: the PrevAKI randomized controlled trial. *Intensive Care Medicine*, *43*(11), 1551–1561.

Mehta, R. H., Grab, J. D., O'Brien, S. M., Bridges, C. R., Gammie, J. S., Haan, C. K., Ferguson, T. B., Peterson, E. D., & Society of Thoracic Surgeons National Cardiac Surgery Database Investigators. (2006). Bedside tool for predicting the risk of postoperative dialysis in patients undergoing cardiac surgery. *Circulation*, *114*(21), 2208–2216; quiz 2208.

Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S. H., Stamm, C., Hofmann, T., Falk, V., & Eickhoff, C. (2018). Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet. Respiratory Medicine*, *6*(12), 905–914.

Mishra, J., Dent, C., Tarabishi, R., Mitsnefes, M. M., Ma, Q., Kelly, C., Ruff, S. M., Zahedi, K., Shao, M., Bean, J., Mori, K., Barasch, J., & Devarajan, P. (2005). Neutrophil gelatinase-associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery. *The Lancet*, *365*(9466), 1231–1238.

Mohamadlou, H., Lynn-Palevsky, A., Barton, C., Chettipally, U., Shieh, L., Calvert, J., Saber, N. R., & Das, R. (2018). Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. *Canadian Journal of Kidney Health and Disease*, *5*, 2054358118776326.

Murty, M. S. N., Sharma, U. K., Pandey, V. B., & Kankare, S. B. (2013). Serum cystatin C as a marker of renal function in detection of early acute kidney injury. *Indian Journal of Nephrology*, *23*(3), 180–183.

Nakayama, M., Ishida, M., Ogihara, M., Hanaoka, K., Tamura, M., Kanai, H., Tonozuka, Y., & Marshall, M. R. (2015). Social functioning and socioeconomic changes after introduction of regular dialysis treatment and impact of dialysis modality: a multi-centre survey of Japanese patients. *Nephrology* , *20*(8), 523–530.

Naughton, C. A. (2008). Drug-induced nephrotoxicity. *American Family Physician*, *78*(6), 743–750.

Neil, D., Pfeiffer, M., & Liu, S.-C. (2016). Phased lstm: Accelerating recurrent network training for long or event-based sequences. *Advances in Neural Information Processing Systems*, 3882–3890.

Nuis, R.-J., Rodés-Cabau, J., Sinning, J.-M., van Garsse, L., Kefer, J., Bosmans, J., Dager, A. E., van Mieghem, N., Urena, M., Nickenig, G., Werner, N., Maessen, J., Astarci, P., Perez, S., Benitez, L. M., Dumont, E., van Domburg, R. T., & de Jaegere, P. P. (2012). Blood transfusion and the risk of acute kidney injury after transcatheter aortic valve implantation. *Circulation. Cardiovascular Interventions*, *5*(5), 680–688.

Pahwa, S., Bernabei, A., Schaff, H., Stulak, J., Greason, K., Pochettino, A., Daly, R., Dearani, J., Bagameri, G., King, K., Viehman, J., & Crestanello, J. (2021). Impact of postoperative complications after cardiac surgery on long-term survival. *Journal of Cardiac Surgery*, *36*(6), 2045–2052.

Palomba, H., de Castro, I., Neto, A. L. C., Lage, S., & Yu, L. (2007). Acute kidney injury prediction following elective cardiac surgery: AKICS Score. *Kidney International*, *72*(5), 624–631.

Parikh, C. R., Coca, S. G., Thiessen-Philbrook, H., Shlipak, M. G., Koyner, J. L., Wang, Z., Edelstein, C. L., Devarajan, P., Patel, U. D., Zappitelli, M., Krawczeski, C. D., Passik, C. S., Swaminathan, M., Garg, A. X., & TRIBE-AKI Consortium. (2011). Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. *Journal of the American Society of Nephrology: JASN*, *22*(9), 1748–1757.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830.

Penny-Dimri, J. C., Bergmeir, C., Reid, C. M., Williams-Spence, J., Cochrane, A. D., & Smith, J. A. (2021). Machine Learning Algorithms for Predicting and Risk Profiling of Cardiac Surgery-Associated Acute Kidney Injury. *Seminars in Thoracic and Cardiovascular Surgery*, *33*(3), 735–745.

Perez, F., & Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. In *Computing in Science & Engineering, 9*(3), pp. 21–29).

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). “Quasi- (that is, Sub-) Random Sequences.” In *Numerical Recipes in FORTRAN: The Art of Scientific Computing* (Vol. 2, pp. 299–306). Cambridge University Press.

Ramesh, G., Krawczeski, C. D., Woo, J. G., Wang, Y., & Devarajan, P. (2010). Urinary netrin-1 is an early predictive biomarker of acute kidney injury after cardiac surgery. *Clinical Journal of the American Society of Nephrology: CJASN*, *5*(3), 395–401.

Rank, N., Pfahringer, B., Kempfert, J., Stamm, C., Kühne, T., Schoenrath, F., Falk, V., Eickhoff, C., & Meyer, A. (2020). Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *NPJ Digital Medicine*, *3*, 139.

Redondo-Pachon, M. D., Enriquez, R., Sirvent, A. E., Millan, I., Romero, A., & Amorós, F. (2014). Acute renal failure and severe thrombocytopenia associated with metamizole. *Saudi Journal of Kidney Diseases and Transplantation: An Official Publication of the Saudi Center for Organ Transplantation, Saudi Arabia*, *25*(1), 121–125.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rosner, M. H., & Okusa, M. D. (2006). Acute kidney injury associated with cardiac surgery. *Clinical Journal of the American Society of Nephrology: CJASN*, *1*(1), 19–32.

Silver, S. A., & Chertow, G. M. (2017). The Economic Consequences of Acute Kidney Injury. *Nephron*, *137*(4), 297–301.

Silver, S. A., Long, J., Zheng, Y., & Chertow, G. M. (2017). Cost of Acute Kidney Injury in Hospitalized Patients. *Journal of Hospital Medicine: An Official Publication of the Society of Hospital Medicine*, *12*(2), 70–76.

Swedko, P. J., Clark, H. D., Paramsothy, K., & Akbari, A. (2003). Serum creatinine is an inadequate screening test for renal failure in elderly patients. *Archives of Internal Medicine*, *163*(3), 356–360.

Thakar, C. V., Arrigain, S., Worley, S., Yared, J.-P., & Paganini, E. P. (2005). A clinical score to predict acute renal failure after cardiac surgery. *Journal of the American Society of Nephrology: JASN*, *16*(1), 162–168.

Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B. B., Rashidi, P., Pardalos, P., Momcilovic, P., & Bihorac, A. (2016). Application of Machine Learning Techniques to High-Dimensional Clinical Data to Forecast Postoperative Complications. *PloS One*, *11*(5), e0155705.

Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C. R., Peterson, K., Reeves, R., Hassabis, D., King, D., Suleyman, M., Back, T., Nielson, C., Ledsam, J.R., & Mohamed, S. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature, 572*(7767),116–119.

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, *13*(2), 22–30.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689.

Wang, Y., & Bellomo, R. (2017). Cardiac surgery-associated acute kidney injury: risk factors, pathophysiology and treatment. *Nature Reviews. Nephrology*, *13*(11), 697–711.

Weyl, H. (1916). Über die Gleichverteilung von Zahlen mod. Eins. *Mathematische Annalen*, *77*(3), 313–352.

Wijeysundera, D. N., Karkouti, K., Dupuis, J.-Y., Rao, V., Chan, C. T., Granton, J. T., & Beattie, W. S. (2007). Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. *JAMA: The Journal of the American Medical Association*, *297*(16), 1801–1809.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution / British Ecological Society*, *3*(1), 129–137.

Ympa, Y. P., Sakr, Y., Reinhart, K., & Vincent, J.-L. (2005). Has mortality from acute renal failure decreased? A systematic review of the literature. *The American Journal of Medicine*, *118*(8), 827–832.

Zaremba, W., Sutskever, I., & Vinyals, O. (2015). Recurrent Neural Network Regularization. In *arXiv [cs.NE]*. arXiv. Last Retrieved October 23, 2021 from http://arxiv.org/abs/1409.2329v5

Eidesstattliche Versicherung / Anteilserklärung

Eidesstattliche Versicherung

„Ich, Nina Rank, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema:

Deep-learning basierte Echtzeit-Vorhersage von akutem Nierenversagen nach kardiochirurgischen Eingriffen (Deep-learning based real-time prediction of acute kidney injury after cardiac surgery)

selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem Erstbetreuer, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Berlin, Unterschrift

Ausführliche Anteilserklärung an der erfolgten Publikation als Top-Journal im Rahmen der Promotionsverfahren zum MD/PhD

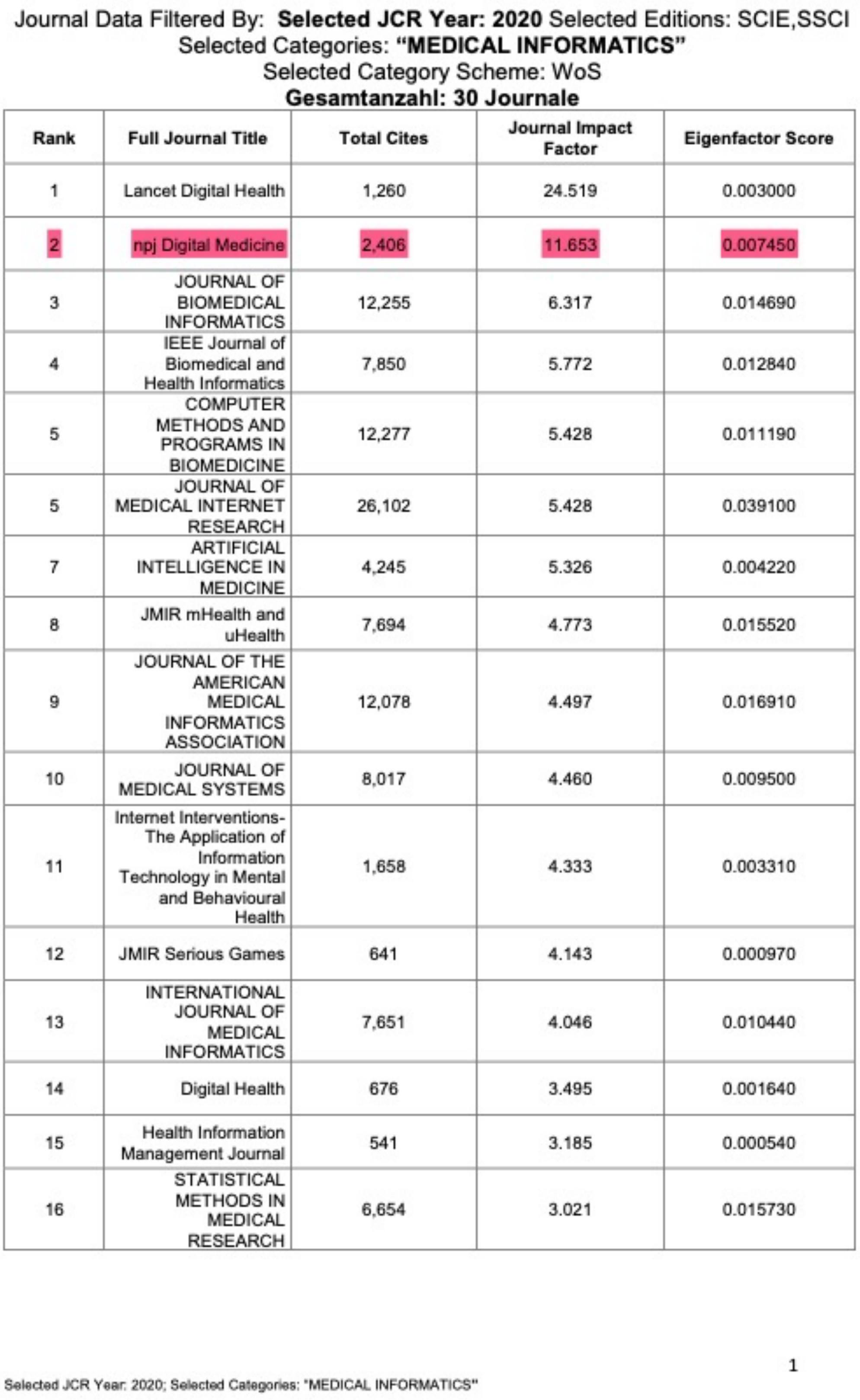
Publikation 1:
Rank Nina, Pfahringer Boris, Kempfert Jörg, Stamm Christof, Kühne Titus, Schoenrath Felix, Falk Volkmar, Eickhoff Carsten, Meyer Alexander. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPJ Digital Medicine. 26.10.2020.

Beitrag im Einzelnen:

Die Studie wurde von Prof. Alexander Meyer und mir in Zusammenarbeit mit mehreren Personen (darunter Prof. Volkmar Falk) konzipiert. Die Methoden wurden von Prof. Alexander Meyer, Prof. Carsten Eickhoff, Boris Pfahringer und mir entworfen. Die Datenakquisition wurde von Prof. Alexander Meyer und mir durchgeführt. Die Vorverarbeitung der Daten und die Entwicklung des RNN erfolgte weitgehend selbstständig durch mich mit technischer Beratung durch Boris Pfahringer. Die manuelle Überarbeitung des unabhängigen Testsets und die Auswertung der Modellleistung wurden von mir durchgeführt. Die Rekrutierung der teilnehmenden Ärzte erfolgte durch Prof. Alexander Meyer und mich. Die statistische Auswertung wurde von mir selbständig durchgeführt. Zu diesem Zweck wurde eine statistische Beratung durch Dr. Konrad Neumann am Institut für Biometrie und Klinische Epidemiologie eingeholt. Alle Tabellen und Abbildungen der Publikation wurden von mir erstellt. Die Erstfassung des Manuskripts wurde von mir verfasst. Es wurde von allen Autoren der Publikation überarbeitet. Die Einreichung des Manuskripts/dessen Revision erfolgte durch mich.

Datum, Unterschrift der Doktorandin

Auszug aus der Journal Summary List



Journal Data Filtered By: **Selected JCR Year: 2020** Selected Editions: SCIE,SSCI
Selected Categories: **“MEDICAL INFORMATICS”**
Selected Category Scheme: WoS
Gesamtanzahl: 30 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	Lancet Digital Health	1,260	24.519	0.003000
2	npj Digital Medicine	2,406	11.653	0.007450
3	JOURNAL OF BIOMEDICAL INFORMATICS	12,255	6.317	0.014690
4	IEEE Journal of Biomedical and Health Informatics	7,850	5.772	0.012840
5	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	12,277	5.428	0.011190
5	JOURNAL OF MEDICAL INTERNET RESEARCH	26,102	5.428	0.039100
7	ARTIFICIAL INTELLIGENCE IN MEDICINE	4,245	5.326	0.004220
8	JMIR mHealth and uHealth	7,694	4.773	0.015520
9	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	12,078	4.497	0.016910
10	JOURNAL OF MEDICAL SYSTEMS	8,017	4.460	0.009500
11	Internet Interventions-The Application of Information Technology in Mental and Behavioural Health	1,658	4.333	0.003310
12	JMIR Serious Games	641	4.143	0.000970
13	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	7,651	4.046	0.010440
14	Digital Health	676	3.495	0.001640
15	Health Information Management Journal	541	3.185	0.000540
16	STATISTICAL METHODS IN MEDICAL RESEARCH	6,654	3.021	0.015730

1

Selected JCR Year: 2020; Selected Categories: "MEDICAL INFORMATICS"

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
17	JMIR Medical Informatics	1,343	2.955	0.003690
18	BMC Medical Informatics and Decision Making	6,015	2.796	0.009140
19	Health Informatics Journal	1,497	2.681	0.002300
20	MEDICAL & BIOLOGICAL ENGINEERING & COMPUTING	7,019	2.602	0.004510
21	MEDICAL DECISION MAKING	6,391	2.583	0.007240
22	Informatics for Health & Social Care	511	2.439	0.000830
23	JOURNAL OF EVALUATION IN CLINICAL PRACTICE	5,408	2.431	0.005340
24	STATISTICS IN MEDICINE	33,374	2.373	0.031200
25	Applied Clinical Informatics	1,481	2.342	0.002900
26	INTERNATIONAL JOURNAL OF TECHNOLOGY ASSESSMENT IN HEALTH CARE	2,522	2.188	0.001760
27	METHODS OF INFORMATION IN MEDICINE	1,601	2.176	0.001340
28	CIN-COMPUTERS INFORMATICS NURSING	1,349	1.985	0.001700
29	Therapeutic Innovation & Regulatory Science	907	1.778	0.002050
30	Biomedical Engineering-Biomedizinische Technik	1,292	1.411	0.001020

Publikation

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance

Nina Rank[1], Boris Pfahringer[1], Jörg Kempfert[1,2], Christof Stamm[1,2], Titus Kühne[2,3,4], Felix Schoenrath[1,2], Volkmar Falk[1,2,4,5,6], Carsten Eickhoff[7] and Alexander Meyer[1,2,4]✉

Acute kidney injury (AKI) is a major complication after cardiothoracic surgery. Early prediction of AKI could prompt preventive measures, but is challenging in the clinical routine. One important reason is that the amount of postoperative data is too massive and too high-dimensional to be effectively processed by the human operator. We therefore sought to develop a deep-learning-based algorithm that is able to predict postoperative AKI prior to the onset of symptoms and complications. Based on 96 routinely collected parameters we built a recurrent neural network (RNN) for real-time prediction of AKI after cardiothoracic surgery. From the data of 15,564 admissions we constructed a balanced training set (2224 admissions) for the development of the RNN. The model was then evaluated on an independent test set (350 admissions) and yielded an area under curve (AUC) (95% confidence interval) of 0.893 (0.862–0.924). We compared the performance of our model against that of experienced clinicians. The RNN significantly outperformed clinicians (AUC = 0.901 vs. 0.745, $p < 0.001$) and was overall well calibrated. This was not the case for the physicians, who systematically underestimated the risk ($p < 0.001$). In conclusion, the RNN was superior to physicians in the prediction of AKI after cardiothoracic surgery. It could potentially be integrated into hospitals' electronic health records for real-time patient monitoring and may help to detect early AKI and hence modify the treatment in perioperative care.

npj Digital Medicine (2020)3:139 ; https://doi.org/10.1038/s41746-020-00346-8

INTRODUCTION

Acute kidney injury (AKI) is a major postoperative complication after cardiothoracic surgery. It is an independent risk factor for early and long-term mortality[1–4] and is strongly associated with increased hospital costs and length of stay[5–7].

AKI is defined as a major increase of serum creatinine or a strong decline in urine output[8]. Compromised renal blood flow and cardiopulmonary bypass play a critical role in the development of AKI, but overall its etiology is highly multifactorial[9–12].

Early detection of patients at high risk of developing AKI allows for early therapeutic intervention prior to the onset of anuria and its complications such as acidosis, hyperkalemia, or volume overload as well as long-term complications such as lung injury, sepsis and chronic kidney disease[13–16]. In a pilot study in 2011 it was demonstrated that in patients with AKI stage I, early nephrologist consultation can avert progression to higher AKI stages[17]. It was also shown that delayed nephrologist involvement (48 h after AKI onset) in critically ill patients was associated with an increase of mortality and dependence on dialysis[18]. An immediate post-operative "KDIGO care bundle" (optimization of volume status and hemodynamics, avoidance of nephrotoxic drugs and hyperglycemia) in high-risk patients has been shown to reduce cardiac surgery-associated AKI[19].

Although several classical clinical risk scores for the prediction of postoperative AKI exist, none of them is specifically recommended by guidelines[20–26]. With few exceptions they rely on patient demographics, disease history and the type of surgery and require time-consuming manual data collection and calculation. Furthermore, they are usually based on static properties or single point-in-time measurements that cannot adapt to the often rapid and dramatic changes that occur in the postoperative setting.

Increased digitization of medical information opens up new alternatives for early prediction of postoperative complications that might potentially be integrated into existing electronic health record (EHR) software. A vast amount of data with high temporal resolution is collected during a hospital stay. Effectively processing such high-dimensional data in a parallelized way, however, goes far beyond the capabilities of the human brain[27]. Machine learning (ML) offers a potential solution to this problem.

Previous studies investigating the performance of ML models in predicting AKI have yielded promising results[28–35]. However, studies directly comparing the predictive performance of ML models against experienced physicians in the prediction of postoperative AKI on time-series data of real clinical cases are highly needed.

We therefore developed a recurrent neural network (RNN) that allows real-time predictions of AKI within the first 7 postoperative days following cardiothoracic surgery based on routinely collected variables (features). This model was then compared to the performance of experienced health-care professionals.

RESULTS

Performance of the RNN based prediction

A complete description of the study population, patient selection process, development of the ML model, and the experimental design of our RNN-vs-human comparison can be found in the 'Methods' section.

[1]Department of Cardiothoracic and Vascular Surgery, German Heart Center Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. [2]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, P.O. Box 65 21 33, 13316 Berlin, Germany. [3]Institute for Computer-assisted Cardiovascular Medicine, Charité–Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. [4]Berlin Institute of Health, Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany. [5]Department of Cardiothoracic Surgery, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. [6]Department of Health Sciences and Technology, ETH Zürich, Leopold-Ruzicka-Weg 4, 8093 Zürich, Switzerland. [7]Center for Biomedical Informatics, Brown University, 233 Richmond Street, Providence, RI 02912, USA. ✉email: meyera@dhzb.de

npj nature partner journals

Table 1. Model performance metrics for balanced test set.

Threshold-independent metrics, (95% CI)			Threshold-dependent metrics, (95% CI)						
AUC	PR_AUC	$\overline{\mathbf{MSE}}_{\mathbf{pat}}$	Acc	Sens	Spec	F1	FPR	NPV	PPV
0.893 (0.862–0.924)	0.903 (0.873–0.933)	0.124 (0.090–0.159)	0.825 (0.786–0.863)	0.853 (0.802–0.904)	0.798 (0.741–0.855)	0.826 (0.776–0.876)	0.202 (0.145–0.259)	0.851 (0.799–0.903)	0.801 (0.745–0.857)

$n = 350$ admissions/patients.
AUC area under curve, *PR_AUC* precision-recall AUC, $\overline{\text{MSE}}_{\text{pat}}$ mean of the brier score of each patient, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity, *F1* F_1-score, *FPR* false-positive rate, *NPV* negative predictive value, *PPV* positive predictive value, *CI* confidence interval. The threshold for positive/negative class prediction was set to 0.41, leading to a sensitivity of 0.850 on cross-validation folds in the training set.

In summary, we retrospectively analysed EHR time series data with high temporal resolution (up to 1 min) generated at a tertiary care center for cardiovascular diseases. Based on $n = 2224$ admissions, we developed an RNN that continuously (every 15 min) predicted the probability of developing AKI defined as KDIGO[8] stage 2 or 3 within the first 7 days after cardiothoracic surgery.

Supplementary Tables 1–4 show a comparison of baseline characteristics between AKI- and non-AKI cases in the training, balanced and imbalanced test set and the whole study population before matching AKI- and non-AKI cases.

Table 1 shows the performance metrics of our RNN evaluated on an independent test set with $n = 350$ patients. The model achieved an area under curve (AUC) (95% confidence interval (CI)) of 0.893 (0.862–0.924). In addition, we trained a model with only serum creatinine as input and yielded an AUC of 0.805 (0.768–0.842). Thus, the addition of further parameters led to an absolute increase of around 10 percentage points in the AUC. However, a model using all features but creatinine and glomerular filtration rate (GFR) (the GFR is calculated from creatinine) performed almost as good as the full model with an AUC of 0.887 (0.855–0.919)—probably due to high correlation between creatinine and other features, e.g., urea. For further performance metrics of these reduced models see Supplementary Tables 5 and 6.

A table with the model performance metrics derived from an imbalanced test set with incidence rate of 10% AKI (see Supplementary Results 1) can be found in Supplementary Table 7. In addition, we analysed some examples of the predictions of individual patients including false-positive and false-negative predictions. These can be found in Supplementary Figs. 1–3.

RNN vs. human-level performance—experimental design

We set up an experiment to compare our ML model against experienced physicians (Fig. 1). For each of the $n = 350$ patients of our balanced test set a quasi-random point in time in their observation period was chosen, further denoted as 'prediction point' (For more information about quasi-random samples see the 'Methods' section.).

At the chosen prediction point, seven experienced physicians and the ML model each had to make a prediction (between 0 and 100%) of how likely the patient was to develop AKI within the first 7 days after surgery.

All time series information up to the 'prediction point' was graphically displayed for the physicians to mimic the electronic patient chart.

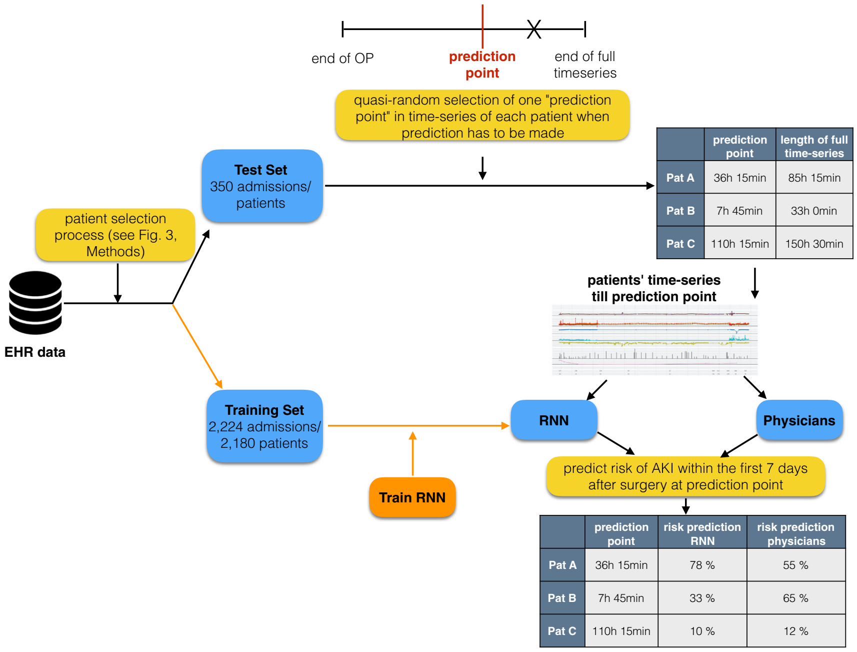
Performance of RNN and physicians

The performance of our RNN and the physicians' assessment can be found in Table 2 (Note that the metrics of the RNN are slightly different from those in section 'Performance of the RNN based prediction'. The reason is that in the RNN vs. human experiment only one prediction point per patient was evaluated, whereas for the complete evaluation of the RNN all predictions of the whole observation periods for all patients were evaluated.).

The median (interquartile range (IQR)) prediction value for the physicians was 0.36 (0.15–0.70) vs. 0.51 (0.12–0.86) for the RNN.

Across all metrics, the RNN outperformed the physicians. We obtained an AUC of 0.901 for the RNN vs. 0.745 for the physicians ($p < 0.001$, $Z = 6.85$, DeLong's test). The receiver operating characteristic (ROC) curves and the precision-recall curves are displayed in Fig. 2a and Fig. 2b, respectively.

The mean of our predictive quality score S ($S = r$, if the patient developed AKI and $S = 1 - r$, if the patient did not develop AKI) was significantly higher for the RNN than for the experienced physicians (0.754 vs 0.639, $p < 0.001$, t-statistic = 8.47, df = 349, paired t-test).


end of OP
prediction point
end of full timeseries
quasi-random selection of one "prediction point" in time-series of each patient when prediction has to be made
Test Set
350 admissions/ patients
patient selection process (see Fig. 3, Methods)
EHR data
patients' time-series till prediction point
Training Set
2,224 admissions/ 2,180 patients
Train RNN
RNN
Physicians
predict risk of AKI within the first 7 days after surgery at prediction point

	prediction point	length of full time-series
Pat A	36h 15min	85h 15min
Pat B	7h 45min	33h 0min
Pat C	110h 15min	150h 30min

	prediction point	risk prediction RNN	risk prediction physicians
Pat A	36h 15min	78 %	55 %
Pat B	7h 45min	33 %	65 %
Pat C	110h 15min	10 %	12 %

Fig. 1 Experimental design for performance comparison of recurrent neural network (RNN) against physicians. The electronic health record (EHR) data was split into a training and a test set. The training set was used for the development of the RNN (orange path). For each patient (Pat) in the test set, a quasi-random 'prediction point' in the time-series was chosen (for more information about quasi-randomness see 'Methods'). EHR data up to this prediction point was given to physicians and RNN (the rest of the time series data, here denoted as X, was hidden). Both physicians and RNN, had to make a prediction for postoperative AKI at this prediction point.

In addition, we investigated the calibration of the RNN's and physicians' predictions. Calibration describes how close the predicted probabilities are to the observed frequencies. A perfectly calibrated model would have one point at (0,0) and one at (1,1) in a calibration plot (it would always predict 0 for negatives and 1 for positives). For a well-calibrated model, the points lie on the diagonal between (0,0) and (1,1). Figure 2c illustrates that in the intervals of high prediction values of physicians, the predicted frequencies of AKI largely correspond to the observed frequencies (upper right part of the calibration curve). However, for several patients that developed AKI, physicians predicted low AKI probabilities (false-negative predictions, lower left part of the calibration curve). This is also reflected in the observation that the physicians' median (IQR) prediction value was lower than the RNN's (Physicians: 0.36 (0.15–0.70) and RNN: 0.51 (0.12–0.86)). Overall the physicians' predictions were not well calibrated ($p < 0.001$, $X^2 = 165.5$, df = 8, Hosmer-Lemeshow-test[36]).

In contrast, Fig. 2d displays a very well calibration ($p = 0.37$, $X^2 = 8.67$, df = 8, Hosmer-Lemeshow-test) for the RNN, with most of the points lying very close to the diagonal, even in intervals of low prediction values.

We investigated the performance of our RNN and physicians at different points in time before the event (AKI or non-AKI/discharge) (see Table 3). Not-surprisingly, both, humans and RNN, performed worse when the event was further away in time. However, low sensitivity rates could also be observed when the event was very close (≤2 h). In this group the median total observation length was very short, meaning that patients who developed AKI, developed it rapidly after surgery. Thus, there was probably not enough information available before the event to reliably predict AKI. However, even in this interval, the RNN reached a sensitivity of 0.789.

DISCUSSION

We developed an RNN for real-time prediction of postoperative AKI within 7 days after cardiothoracic surgery—based on routinely collected features during the hospital stay and then retrospectively validated it on an independent test set.

To test the clinical significance, we performed a side-by-side comparison of our model against experienced physicians. Such direct comparisons are highly needed, but hardly ever performed in clinical ML studies. We had expected our model to perform nearly as well as the physicians, and had designed our study as a non-inferiority-experiment. Surprisingly, our RNN significantly outperformed experienced clinicians in terms of the mean of our performance metric S. (S indicates how close a prediction is to the observed outcome). In addition, the model reached a significantly higher AUC than the physicians (0.901 vs. 0.745, $p < 0.001$, DeLong's test) and was overall well calibrated (Hosmer-Lemeshow-Test: $p = 0.37$ vs. $p < 0.001$ for physicians).

Physicians showed an overall low sensitivity of 0.594 at AKI prediction. They predicted lower risk probabilities in general. They reached a maximum sensitivity of 0.793 for the 2–6 h interval before the event and a minimum sensitivity of 0.387 for the

Table 2. Performance metrics of recurrent neural network (RNN) and physicians on a balanced test set.

	Threshold-independent metrics, (95% CI)			Metrics based on a threshold of 0.5 for positive/negative classification, (95% CI)						
	AUC	PR_AUC	Brier	Acc	Sens	Spec	F1	FPR	NPV	PPV
RNN	0.901 (0.870–0.932)	0.907 (0.877–0.937)	0.122 (0.088–0.156)	0.846 (0.808–0.884)	0.851 (0.798–0.904)	0.840 (0.787–0.894)	0.847 (0.797–0.897)	0.160 (0.106–0.214)	0.850 (0.797–0.903)	0.842 (0.788–0.896)
Physicians	0.745 (0.699–0.791)	0.747 (0.701–0.793)	0.217 (0.174–0.260)	0.711 (0.664–0.759)	0.594 (0.521–0.667)	0.829 (0.773–0.884)	0.673 (0.609–0.738)	0.171 (0.116–0.227)	0.671 (0.601–0.741)	0.776 (0.715–0.838)

n = 350 admissions/patients.
AUC area under curve, *PR_AUC* precision-recall AUC, *Brier* Brier score, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity, *F1* F_1-score, *FPR* false-positive rate, *NPV* negative predictive value, *PPV* positive predictive value, *CI* confidence interval.

24–48 h interval before the event. Thus, they systematically underestimated the risk of AKI. This suggests that physicians mainly recognize AKI stage 3 or dialysis and that lower AKI stages are erroneously considered unproblematic. It has been demonstrated, however, that even minor increases in serum creatinine after cardiac surgery are associated with an increased mortality risk[37].

The participating physicians each had at least one year working experience on a cardiothoracic intensive care unit (ICU), but were no specialists in nephrology. This reflects a realistic clinical setting on an ICU, where nephrologists are usually not available around the clock.

In contrast to the physicians, our RNN yielded an overall high sensitivity of 0.851 with a maximum sensitivity of 0.971 in the 2–6 h interval before the event and a minimum sensitivity of even 0.750 in the 48–168 h interval before the event. In summary, our RNN was superior to experienced physicians in the prediction of AKI after cardiothoracic surgery.

From a modeling point of view, our RNN could easily be integrated into an EHR system. It does not require any additional human input as all data transformation is implemented programmatically. Allowing for personalized predictions, it may enable earlier identification and intervention in high-risk patients and thus contribute to an improvement of patient care and safety. However, the transfer of such a retrospective model from research to real implementation raises additional challenges. Technical barriers, data security when exporting personal data to external software systems, and business considerations may be diverse and can conflict with each other.

Our model achieved highly accurate results with an overall AUC of 0.893 in our internal validation. It outperformed existing classical prediction models that are based on logistic regression from static pre- and intraoperative variables, as well as a dynamic model that predicted AKI at three points in time (pre-operative, at ICU admittance and 24 h after ICU admittance). These models reached AUCs ranging from 0.72–0.85 in their respective internal validation cohorts and used slightly different definitions of AKI[20–26,38] (see Table 4). The proposed model does not create additional workload for physicians, as it only used routinely collected data of the EHR. As such, it only employs data that is available at the time of prediction and all data transformations are implemented programmatically. It is worth noting that the model performed very well, although it was built on a relatively small sample size of 2224 admissions.

Previous studies have demonstrated the benefits of using ML for AKI prediction. Thottakkara et al.[28] applied different ML approaches to forecast postoperative AKI and observed promising performances in their internal validation cohort (AUC between 0.797 and 0.858). Bihorac et al.[29] used an ML algorithm to assess the risk of 8 postoperative complications including AKI and reported an AUC of 0.80 (0.79–0.80) for AKI prediction. The approach of both studies, however, relied exclusively on static, mostly preoperative features.

A multi-center ward-based AKI prediction model was developed by Koyner et al.[39] using a discrete time survival model with an AUC (95% CI) of 0.76 (0.76–0.77) for AKI of at least stage 2.

In 2018, Koyner et al.[31] published another study using EHR data for AKI risk prediction and reached an AUC (95% CI) of 0.90 (0.90–0.90) for predicting stage 2 AKI within the next 24 h and 0.87 (0.87–0.87) within the next 48 h. Cheng et al.[32] built ML models to forecast AKI over various time horizons and obtained an AUC of 0.765 (prediction one day before the event). In these studies, however, the urine output criterion of AKI, a central component in the KDIGO definition was not integrated, which can lead to a false-negative classification of AKI cases. In our training and test cohort around 30% of the AKI cases were defined by the urine criteria of KDIGO (see Supplementary Table 8). We can assume that a substantial proportion of the patients in the above studies would

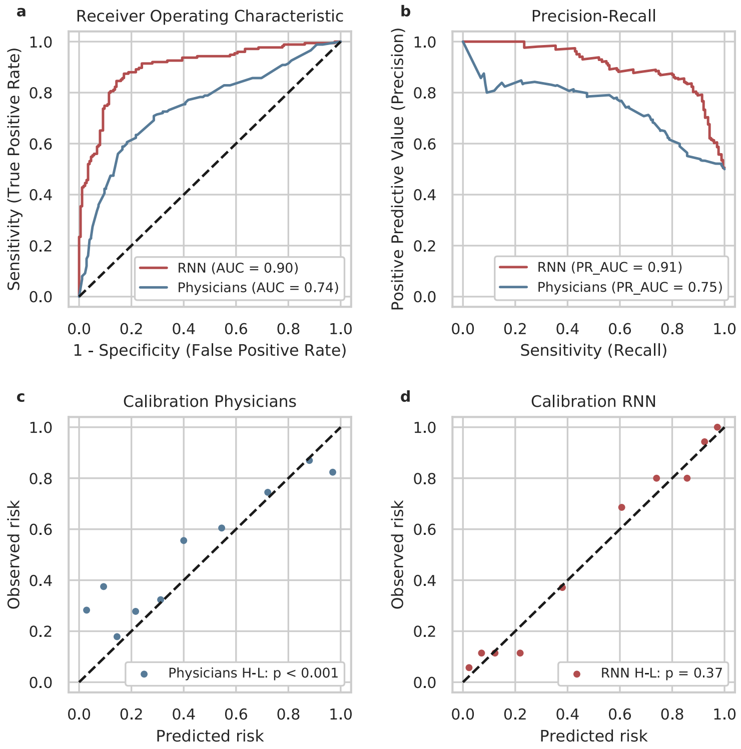


Fig. 2 Discrimination and calibration of the predictions of recurrent neural network (RNN) and physicians. a receiver operating characteristics (ROC), **b** precision-recall curve**, c** calibration of physicians, **d** calibration of RNN. AUC area under curve. H-L Hosmer-Lemeshow-Test[36], PR_AUC precision-recall AUC. The RNN outperformed clinical physicians regarding AUC (**a**) and PR_AUC (**b**). Physicians systematically underestimated the risk of acute kidney injury (predicted risks < observed risks, **c**). In contrast, the RNN was overall well calibrated (**d**).

Table 3. Performance metrics of recurrent neural network (RNN) and physicians in temporal dependence to the event.

Predictor	Time to event	patients	AKI	MOL	AUC	PR_AUC	Brier	Acc	Sens	Spec	F_1	FPR	NPV	PPV
RNN	0 h to 2 h	54	19	8.3 h	0.913	0.837	0.113	0.870	0.789	0.914	0.811	0.086	0.889	0.833
Physicians	0 h to 2 h	54	19	8.3 h	0.709	0.552	0.199	0.759	0.632	0.829	0.649	0.171	0.806	0.667
RNN	2 h to 6 h	63	29	12.5 h	0.881	0.88	0.13	0.825	0.862	0.794	0.820	0.206	0.871	0.781
Physicians	2 h to 6 h	63	29	12.5 h	0.853	0.861	0.152	0.794	0.793	0.794	0.780	0.206	0.818	0.767
RNN	6 h to 12 h	63	34	17.8 h	0.942	0.948	0.088	0.921	0.971	0.862	0.930	0.138	0.962	0.892
Physicians	6 h to 12 h	63	34	17.8 h	0.811	0.798	0.19	0.746	0.618	0.897	0.724	0.103	0.667	0.875
RNN	12 h to 24 h	74	42	36.4 h	0.888	0.921	0.128	0.824	0.881	0.750	0.851	0.250	0.828	0.822
Physicians	12 h to 24 h	74	42	36.4 h	0.693	0.706	0.257	0.689	0.667	0.719	0.709	0.281	0.622	0.757
RNN	24 h to 48 h	60	31	46.4 h	0.890	0.899	0.142	0.817	0.774	0.862	0.814	0.138	0.781	0.857
Physicians	24 h to 48 h	60	31	46.4 h	0.718	0.774	0.246	0.633	0.387	0.897	0.522	0.103	0.578	0.800
RNN	48 h to 168 h	36	20	99.0 h	0.875	0.929	0.132	0.806	0.750	0.875	0.811	0.125	0.737	0.882
Physicians	48 h to 168 h	36	20	99.0 h	0.647	0.741	0.274	0.611	0.400	0.875	0.533	0.125	0.538	0.800

AKI number of patients with acute kidney injury, *MOL* median total observation length, *AUC* area under curve, *PR_AUC* precision-recall AUC, *Brier* Brier score, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity, *F1* F_1-score, *FPR* false-positive rate, *NPV* negative predictive value, *PPV* positive predictive value.

Table 4. Comparison between classical prediction models[20] based on logistic regression and our recurrent neural network (RNN).

Authors, model	Sample size derivation	Sample size internal validation	Validation method	"Real-time" prediction	Predicted outcome	Manual calculation	AUC on internal validation
Chertow et al., CICSS[21]	42,773	42,773	100-sample bootstrap	No	30 days post-op. AKI	Yes	0.76 (AUC on derivation cohort)
		3795	Prospective validation				Not reported
Brown et al., NNECDSG[38]	8363	8363	Bootstrap validated C-index (AUC)	No	Severe post-op. AKI (eGFR < 30 ml/min)	Yes	0.72* (0.68–0.75)
Palomba et al., AKICS[24]	603	215	Prospective validation	No	7 days post-op. AKI	Yes	0.85 (0.8–0.9)
Aronson et al., MCSPI[25]	2381	2420	Split sample validation	No	Renal dysfunction or renal failure (dialysis or evidence of renal failure at autopsy)	Yes	0.80
Wijeysundera et al., SRI[26]	10,751	10,751	200-sample bootstrap	No	Post-op. renal replacement therapy	Yes	0.81* (0.78–0.84)
		2566	Prospective validation				0.78 (0.72–0,84)
Mehta et al., STS (Mehta)[23] simplified model	449,524	86,009	Independent sample	No	Post-op. dialysis	Yes	0.83
Thakar et al., Cleveland Clinic[22]	15,838	15,839	Split sample validation	No	Post-op. dialysis	Yes	0.82 (0.80–0.85)
Jiang et al., Dynamic Predictive Score[67]	6081	1152	Independent sample	No	AKI ≥ stage 1 KDIGO	Yes	0.74 preoperative, 0.75 at ICU admission, 0.82 postoperative
This study, RNN	2224	350	Independent Sample (balanced, incidence 50%)	Yes	7 days post-op. AKI stage 2 or 3	No	0.89 (0.86–0.92)
		1945	Independent sample (imbalanced, incidence 10%)				0.85 (0.83–0.86)

AKI acute kidney injury, *AUC* area under curve.

also have met the urine criteria first. Probably not all of them have been classified as false-negative, as they might have met the creatinine criterion at a later stage. In our population, 11% of the AKI-cases in the training set and 12% in the test set exclusively fulfilled the urine criterion and would have been diagnosed false-negatively without this criterion. The median (IQR) diagnosis delay of patients who met both criteria within 7 postoperative days was 14.0 h (6.3–27.3 h) in the training set and 13.3 h (5.3–22.4 h) in the test set. Especially in models with short prediction horizons, there is a high risk that the prediction of imminent AKI and consequently initiation of preventive measurements is delayed when not integrating the urine criterion.

In addition, these previous models were restricted to patients with a serum creatinine of <3 mg/dl (Koyner et al.) or even normal serum creatinine level and a GFR of at least 60 ml/min/1.73 m^2 (Cheng et al.) at admission.

Mohamadlou et al.[40] developed an ML algorithm based on EHR data for detection of AKI at onset and prediction of AKI 12, 24, 48, and 72 h before onset. They reported AUCs from 0.872 (onset) to 0.728 (72 h before onset).

Another study for continuous AKI prediction on a large data set was performed by Tomašev et al.[34]. The developed RNN predicted AKI stage 2 or 3 with an AUC of 0.971 24 h before onset.

Also in these studies the urine output criterion of AKI was not incorporated. In addition, in the study of Tomašev et al. only patients were included for whom at least one year of EHR data were available before admission. They added aggregate features of up to five years of historical information of each individual patient. This approach requires that patients are already known in the admitting hospital, which is often not the case. It is unclear how their algorithm would perform on patients without any prior medical history. In contrast, we used a real uncurated data stream in our model that only contained information generated after admission.

Meyer et al.[35] used an RNN to predict AKI requiring dialysis, mortality and postoperative bleeding after cardiac surgery using routinely collected parameters within the first 24 hours after surgery. The deep-learning model provided very accurate predictions (positive predictive value (PPV)/sensitivity for AKI: 0.87/0.94) that outperformed usual clinical risk scores.

Our model predicted AKI in a time frame up to 7 days after cardiothoracic surgery. Compared to the observation windows of the studies mentioned above, this is a much longer time period. Events in the near future are usually easier to predict than those in the more distant future. To intervene early when the kidneys are merely at risk of injury, a longer prediction window might be necessary. It has been shown that early intervention can prevent AKI or its progression to higher stages[17,19]. Therefore, the prediction of our model was not limited to AKI requiring dialysis,

X

but included the prediction of AKI stages 2 or 3 according to the KDIGO definition.

To conclude, based on a relatively small sample size, we developed a highly accurate model for the prediction of AKI after cardiac surgery that significantly outperformed experienced physicians, could potentially be integrated into EHR systems and might prevent severe complications following AKI through real-time patient surveillance. In a long-term perspective, an extension of the application from a simple risk prediction model to treatment decision support tool is also conceivable.

This study has several shortcomings. The observation periods of the included patients varied widely in length. For most patients it ended in <3 days while some outliers lasted for up to 7 days. We only used the start of nephrotoxic drug administration as a feature. Consideration of exact dose, administration route (e.g., i.v., p.o, …), and administration length could reflect the underlying pharmacodynamics better and improve the prognostic performance.

Our RNN is currently cohort specific for cardiothoracic surgery patients that most likely have different characteristics and risk factors than, e.g., neurosurgical patients. Implementing the same approach on other patient cohorts could give a deeper insight into the generalizability of our method.

Our study is retrospective. Thus, in our RNN vs. physicians head-to-head comparison, physicians only received EHR data and could not clinically evaluate patients. Information such as volume status (except for weight), general condition, etc. or additional examinations (e.g., ultrasound) were not available to them and to the RNN. This deviation from the physicians' usual workflow in clinical practice may explain some of the observed performance deficits. Real clinical data can be very noisy, leading to reduced performance and greater burden of deploying completely automated systems. This stresses once again the fact that artificial intelligence should be utilised in support systems for physicians and not as their replacement.

External validation trials should be performed on prospective data. In addition, they should focus on usage and acceptance of a system such as the one described here in a real clinical setting.

METHODS

Ethics and reporting guideline

This study was approved by the institutional data protection officer and ethics committee of Charité – Universitätsmedizin Berlin (EA2/180/17). The approval included the collection of data on implied consent. We only used retrospective data and the patients were not actively involved in the study. The requirement of informed consent of the participating physicians was waived by the Institutional Review Board (IRB) of Charité – Universitätsmedizin Berlin due to anonymized data acquisition. Reporting of development and validation of the prediction model follows widely the guideline of the TRIPOD statement[41].

Patient selection process

We retrospectively analysed EHR time series data generated between October 2012 and February 2018 at a tertiary care center for cardiovascular diseases.

We included adult patients (18+) that were admitted at least once to the operating theatre for cardiothoracic surgery (15,564 admissions/13,895 patients). We excluded patients without any creatinine or urine flow values, patients receiving hemodialysis before the end of the operation or having a baseline creatinine level ≥4.0 mg/dl (2322 admissions/1487 patients).

Within this collection of 12,978 admissions, 1308 cases were identified with severe postoperative AKI defined as stage 2 or 3 according to KDIGO AKI guidelines—briefly, an increase in serum creatinine to at least twice the baseline value or a decrease in urine flow < 0.5 ml/kg/h for ≥12 h.

As AKI can develop over multiple days, we defined a study period of 7 days after cardiothoracic surgery. The global AKI label of a patient was set positive when the KDIGO criteria stage 2 or 3 was fulfilled at any point within these 7 postoperative days.

The observation time of each patient started when the patient was transferred to the ICU or recovery room. It ended when the patient was either discharged, or when the KDIGO criteria for AKI stage 2 or 3 were fulfilled, or after 7 days after the end of the first surgery.

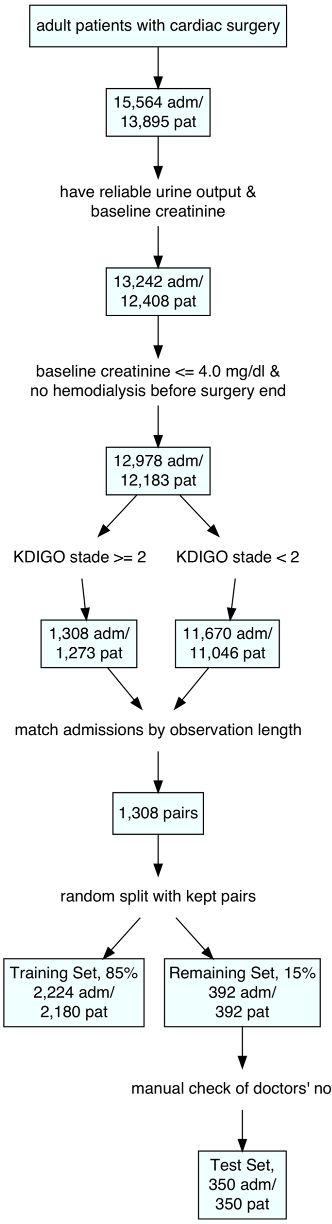


Fig. 3 Flow chart of patient selection process. adm admissions, pat patients.

Each AKI-case was assigned a control out of the non-AKI pool (11,670 admissions/11,046 patients). The controls were matched to the cases on

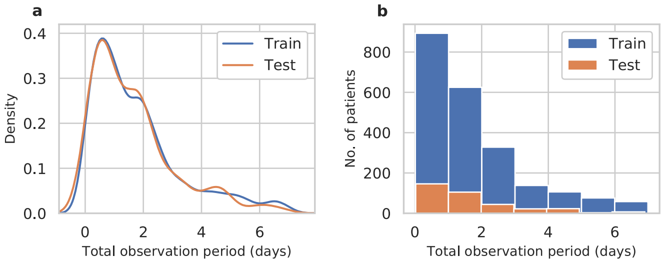


Fig. 4 Total observation period for the training and test set. a Density distribution. **b** Histogram. For most patients the observation period ended within three days after surgery.

observation length. Thus, we generated a balanced data set that we then randomly split into a training set (85%, 2224 admissions/2180 patients) and the remaining set (15%, 392 admissions/patients) while keeping the cases with their respective controls.

For the 392 patients of the remaining set we manually checked physicians' notes in the EHR data and consequently excluded 28 patients. Exclusion criteria were primarily insufficient documentation of the type of surgery, false recording of surgery times or notion of end-stage kidney disease in the patients' history that was not detected by automated filtering.

Out of this set, we randomly selected 350 patients that formed the final test set for model evaluation and comparison with human-level performance. A detailed flow chart of the patient selection process is shown in Fig. 3.

The baseline characteristics were well balanced between the training and the test and are summarized in Supplementary Table 8.

The density distribution and a histogram of the observation periods for patients in the training and test sets is shown in Fig. 4. Most patients were either discharged or diagnosed with AKI within the first 3 days after the first surgery.

Feature selection and preprocessing

We developed our model based on 96 routinely collected clinical parameters. Table 5 gives an overview of all considered features. They can be grouped into static features (e.g., most patient and surgery characteristics, 25 features) that do not change over the observation period and frequently measured dynamic features (e.g., lab values, vital signs, blood gas values and fluid output, 49 features). In addition, we included a variety of widely administered agents that have been reported to potentially cause nephrotoxic effects[42–47] (22 features).

The last creatinine/urea value before surgery was used as a baseline. If there was none available in the five days before surgery, we used the first postoperative value.

We observed that urine output was sometimes incompletely documented on normal wards. As this could lead to false-positive AKI diagnoses we considered urine values reliable only when they were recorded in the operation theatre, the recovery room or the ICU. Thus, on normal wards AKI was only defined by the creatinine criterion whereas in the recovery room or the ICU both AKI criteria (creatinine and urine) were used.

EHR systems are often designed with billing and revision purposes in mind, making certain retrospective therapeutic analyses difficult to conduct due to missing information[48]. In our case, the type of operation that patients underwent was available partly in unstructured textual and partly in categorical form. To access both types of data, we developed a separate set of bag-of-words logistic regression models that predicted the type of operation based on unstructured text describing the operation procedures. As explanatory variables we used all single words or abbreviations that occurred in the pool of text information in its training set. The probability of a specific surgery type Y_i ($i = 1, 2, \ldots, 17$) was given by

$$P(Y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)} \tag{1}$$

where x_j, denotes a count variable indicating how often word/abbreviation j occurred in a patient's surgery procedure description ($j = 1, 2, \ldots,$ no. distinct words/abbreviations). For further information see Supplementary Note 1, Supplementary Tables 9 and 10.

Time sequences with 15-min intervals of all features served as input to our model.

Except for the nephrotoxic agents, missing values were filled by forward imputation. If no precedent value was available, static default values defined by a clinical expert were imputed (one value per feature). The same default values were used for all patients and they were imputed programmatically. They are shown in Supplementary Table 11.

It is extremely difficult to determine the exact effect duration of a drug due to varying excipients, dosages, drug combinations, application types and patient conditions. Therefore, the administration of a drug was considered as an event. For each nephrotoxic agent class in Table 5 a binary feature was created and its value was set to 1 only at the single time slice immediately following the administration of the drug.

Except for the operation types all continuous features were then scaled as follows[49]:

$$X_{\text{scaled}} = \frac{X - \mu(X_{\text{train}})}{IQR(X_{\text{train}})} \tag{2}$$

where $\mu(X_{\text{train}})$ denotes the median and $IQR(X_{\text{train}})$ the IQR of the feature X in the training set. In total, the model was built on a data matrix of 36,244,608 single data points.

For patient selection, preprocessing of features and imputation of missing data, we used R v3.3.3 (R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/) and Python v3.6.7 (The Python Software Foundation, Beaverton, OR) with modules IPython[50] (v7.5.0), Matplotlib[51] (v3.1.0), Scikit-learn[52] (v0.19.1), Pandas[53] (v0.24.2) and Numpy[54] (v1.16.2).

Modeling

In contrast to classical prediction models such as logistic regression, RNNs are able to capture the temporal development of features in a truly sequential fashion as they incorporate information about preceding time steps, links between single timesteps and a direct indicator of the current position in the timeline (see Fig. 5).

We constructed a set of RNNs with different architectures (preceding convolutional layer, different cell types) which allow to process dynamic temporal information.

Hyperparameter tuning was performed on the training set using fivefold cross-validation with balanced class proportions in each fold. We used the Adam optimizer[55] with a fixed learning rate of 0.001. The hyperparameter configurations leading to the highest overall AUC on cross-validation folds of the training set were chosen as final models.

As the parameters of an RNN depend on their initialization and the order in which the training instances are presented, 10 final models with the same hyperparameters but different initializations were trained on the training set. Our final model comprised a uniform ensemble of the 10 constituent models.

For the modeling process we used Python v3.6.7 (The Python Software Foundation, Beaverton, OR) with modules Tensorflow[56], IPython[50] (v7.5.0), Matplotlib[51] (v3.1.0), Scikit-learn[52] (v0.19.1), Pandas[53] (v0.24.2) and Numpy[54] (v1.16.2).

Table 5. Input feature overview.

Feature Group (no. features)	Features
Patient characteristics (4)	Age, sex, weight, height
Laboratory results (25)	Phosphate, total bilirubin, baseline creatinine, creatinine, baseline urea, urea, glomerular filtration rate, creatine kinase (CK), CK-MB, red blood count, white blood count, platelets, C-reactive protein, gamma-glutamyltransferase, glutamic oxaloacetic transaminase, hemoglobin, international normalized ratio, lactate dehydrogenase, magnesium, hematocrit, prothrombin time, partial thromboplastin time, mean corpuscular hemoglobin, mean corpuscular volume, mean corpuscular hemoglobin concentration
Surgery characteristics (20)	Aortic cross-clamp time, cardiopulmonary bypass time, time in operation theatre, surgery procedure (from logistic regression text model, see Supplementary Note 1)
Vital signs (8)	Systolic, mean and diastolic arterial pressure, central venous pressure, heart frequency, pulse, body temperature, oxygen saturation
Arterial blood gas values (BGA) (15)	Base excess, bicarbonate, glucose, hemoglobin, oxygen saturation, partial pressure of carbon dioxide and oxygen, total carbon dioxide, pH level, potassium, sodium, calcium, lactate, carboxyhemoglobin, oxyhemoglobin
Fluid output (2)	Bleeding Rate, urine flow rate
Nephrotoxic agents (22)	Allopurinol, Aminoglycosides, Amphotericin B, Antiplatelet agents (clopidogrel, ticlopidine), Benzodiazepines, Cephalosporins, Cyclosporine, Haloperidol, Ketamine, Nonsteroidal anti-inflammatory drugs, Paracetamol, Penicillines, Proton pump inhibitors, Pyrazolone derivatives, Quinolones, Ranitidine, Rifampin, Sulfonamides, Tacrolimus, (Val-)/Ganciclovir, Aciclovir, Vancomycin Red Blood Cell Transfusions

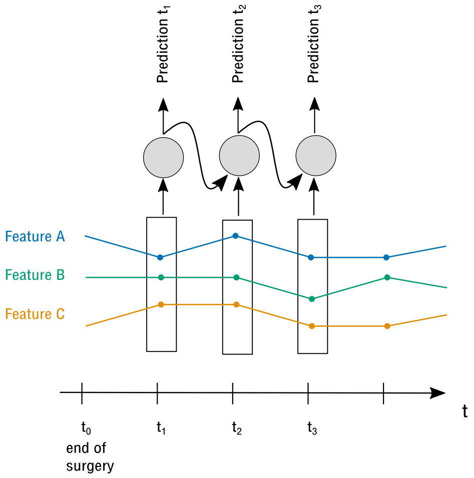


Fig. 5 Architecture of a recurrent neural network (RNN). At each time step, the model receives the current time slice data as input as well as the own output from the preceding time step. The features are captured in a truly sequential fashion.

Measuring RNN performance

We measured the performance of the RNN on an independent test set. No instance of this test set was used for training of the final model. We calculated AUC, precision-recall-AUC (PR_AUC), accuracy, sensitivity, specificity, PPV, negative predictive value (NPV), false-positive rate (FPR) and the F_1-score to measure prediction correctness.

In addition, we calculated the mean of the Brier score[57]—or mean squared error—of each patient ($\overline{\text{MSE}}_{\text{pat}}$)—a measure of accuracy of predictions, without the need for a set threshold.

A single patient's Brier score—or mean squared error—is calculated as follows:

$$\text{MSE}_{\text{pat}} = 1/ts_j \sum_{i=0}^{ts_j} (y_{ji} - y_{jt})^2 \tag{3}$$

where ts_j is the number of timesteps, y_{ji} the prediction at time step i and y_{jt} the true label of patient j.

The $\overline{\text{MSE}}_{\text{pat}}$ ranges from 0 to 1, with value 0 meaning perfect prediction and 1 meaning worst prediction. Random guessing (always predicting 50%) would result in a $\overline{\text{MSE}}_{\text{pat}}$ of 0.25. In contrast to the metrics mentioned above, the $\overline{\text{MSE}}_{\text{pat}}$ is independent of the individual observation length of a patient and the resulting number of predictions per patient.

We adjusted the threshold for positive class prediction until a fixed sensitivity of 0.85 on cross-validation folds in the training set was reached (threshold = 0.41).

Our model predicted the risk of developing AKI every 15 min after the initial surgery. The predictions of an individual patient can be regarded as a cluster of usually highly correlated data. We therefore had to adjust the CIs of our model's metrics. We calculated the 95% CI of each metric X as follows:

$$X + -1.96\sigma(X)$$

with a standard error $\sigma(X)$ of variable X of

$$\sigma(X) = \sqrt{\frac{X(1-X)}{n_{eff}}} \tag{4}$$

To account for intracluster correlation, our sample size n was adjusted, resulting in an effective sample size of[58,59]

$$n_{\text{eff}} = \frac{n}{\text{DE}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m_i} 1}{\text{DE}} \tag{5}$$

where k is the number of patients and m_i the number of time steps of patient i. DE denotes the design effect, also called variance inflation factor, and can be calculated as follows[60]:

$$\text{DE} = \frac{\overline{m}k}{\sum_{i=1}^{k} \frac{m_i}{1+(m_i-1)\text{ICC}}} \tag{6}$$

with ICC as the intracluster correlation coefficient. The ICC was calculated using the R package ICC[61] (v2.3.0).

Comparing RNN vs. human performance

We set up an experiment to compare the performance of our RNN against that of experienced physicians (see Fig. 1). For each patient in the test set, a quasi-random point in time in their observation period was chosen, further denoted as the 'prediction point'. In contrast to real uniform random samples, which tend to form clusters and contain regions without any points at all, quasi-random sequences reduce the probability of cluster formation while still being uniformly distributed[62,63]. This method prevented us from accidentally exclusively sampling prediction points from e.g. the first half of the patients' observation periods.

At each prediction point, a physician and the RNN had to predict whether a patient would develop AKI within the first 7 days after surgery.

All time series information up to the 'prediction point' was graphically displayed for the physicians to mimic the electronic patient chart—although here not in 15-min intervals but in the originally recorded time resolution (up to 1 min).

To create a realistic setting, physicians not only received information about nephrotoxic agents, but of all administered drugs. In addition, the surgery type was given to them as unstructured text manually extracted from physicians' notes. This information was not available to the RNN model. Physicians were explicitly informed about the incidence rate of 50% AKI in our test set.

A physician as well as the RNN made a probability prediction r of the development of AKI for each patient at the respective prediction point. In addition, the physicians made a binary decision (development of AKI: yes/no).

We asked 14 physicians to participate in our study, 10 of whom agreed (response rate = 0.71). All had to meet the selection criteria of ≥5 years of clinical experience and ≥1 year of work experience on a cardiothoracic ICU. From the 10 volunteers we selected seven physicians with different levels of expertise (senior resident up to senior consultant) to create a most realistic setting. Their working experience on a cardiothoracic ICU ranged from at least one year up to several years. None of the participating physicians were specialists in nephrology as nephrologists are usually not constantly available on an ICU. Each physician made predictions for 50 different patients.

Statistical analysis

The initial aim of our study was to show that the RNN is not inferior to experienced physicians in the prediction of AKI. For both, RNN and physicians, the predictive quality of each probability prediction r was measured by a score S as follows:

$$S = r, \text{if the patient developed AKI}$$

$$S = 1 - r, \text{if the patient did not develop AKI}$$

A prior investigation of the RNN's predictions had shown that S was non-normally distributed. Thus, for sample size calculation and power analysis we considered the transformed score X, which was approximately normally distributed:

$$X = -\log(-\log(S)) \quad (7)$$

We assumed that X of the physicians' predictions would also be normally distributed.

Based on a significance level of $\alpha = 0.025$, a power of at least 80% and a non-inferiority margin of $\delta = 0.3$ (this corresponds to a non-inferiority margin of 5.5% for sensitivity + specificity), we obtained a sample size of $N = 350$.

Both, for RNN and physicians, we calculated AUC, PR_AUC, brier score, accuracy, sensitivity, specificity, PPV, NPV, FPR and F_1-score. We set the threshold for positive class prediction to 0.5 as this was also the threshold in the physicians' predictions that corresponded to the 'yes/no'-classification. We calculated CIs for all metrics as described in Section 'Measuring RNN Performance' whereas the effective sample size was $n_{\text{eff}} = n = 350$ as there was no clustering.

For the statistical comparison of S between RNN and physicians we applied a paired t-test. We used DeLong's[64] method to compare the two correlated ROC curves using the R package pROC[65] (v1.9.1). In addition, we investigated the calibration of both, physicians' and RNN's predictions, with the Hosmer-Lemeshow-Test using the R package ResourceSelection[66] (v0.3-2). All three comparisons mentioned above were tested on a significance level of $\alpha = 0.05$.

Reporting summary

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

DATA AVAILABILITY

The EHR data used in this study contain protected health information (PHI) and cannot be published for reasons of data protection. The dataset may be available from the German Heart Center Berlin subject to ethical approvals.

CODE AVAILABILITY

The source code of this publication may be obtained from the corresponding author upon request. To enable independent replication, we describe the experiments and implementation in detail in the 'Methods' section and in the Supplementary Information.

Received: 10 February 2020; Accepted: 17 September 2020;
Published online: 26 October 2020

REFERENCES

1. Chertow, G. M., Levy, E. M., Hammermeister, K. E., Grover, F. & Daley, J. Independent association between acute renal failure and mortality following cardiac surgery 12. *Am. J. Med.* **104**, 343–348 (1998).
2. Hobson, C. E. et al. Acute kidney injury is associated with increased long-term mortality after cardiothoracic surgery. *Circulation* **119**, 2444–2453 (2009).
3. Mandelbaum, T. et al. Outcome of critically ill patients with acute kidney injury using the Acute Kidney Injury Network criteria. *Crit. Care Med.* **39**, 2659–2664 (2011).
4. Ympa, Y. P., Sakr, Y., Reinhart, K. & Vincent, J.-L. Has mortality from acute renal failure decreased? A systematic review of the literature. *Am. J. Med.* **118**, 827–832 (2005).
5. Hobson, C. et al. Cost and mortality associated with postoperative acute kidney injury. *Ann. Surg.* **261**, 1207–1214 (2015).
6. Silver, S. A., Long, J., Zheng, Y. & Chertow, G. M. Cost of acute kidney injury in hospitalized patients. *J. Hosp. Med.* **12**, 70–76 (2017).
7. Silver, S. A. & Chertow, G. M. The economic consequences of acute kidney injury. *Nephron* **137**, 297–301 (2017).
8. Khwaja, A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin. Pract.* **120**, c179–c184 (2012).
9. Spanuchart, I., Cheungpasitporn, W., Thongprayoon, C., Ratanapo, S. & Srivali, N. Off-pump versus on-pump coronary artery bypass surgery: an updated meta-analysis of randomized controlled trials on acute kidney injury and mortality outcomes. *J. Am. Coll. Cardiol.* **65**, A211 (2015).
10. Seabra, V. F., Alobaidi, S., Balk, E. M., Poon, A. H. & Jaber, B. L. Off-pump coronary artery bypass surgery and acute kidney injury: a meta-analysis of randomized controlled trials. *Clin. J. Am. Soc. Nephrol.* **5**, 1734–1744 (2010).
11. Mao, H. et al. Cardiac surgery-associated acute kidney injury. *Blood Purif.* **37**(Suppl 2), 34–50 (2014).
12. Wang, Y. & Bellomo, R. Cardiac surgery-associated acute kidney injury: risk factors, pathophysiology and treatment. *Nat. Rev. Nephrol.* **13**, 697–711 (2017).
13. Faubel, S. & Shah, P. B. Immediate consequences of acute kidney injury: the impact of traditional and nontraditional complications on mortality in acute kidney injury. *Adv. Chronic Kidney Dis.* **23**, 179–185 (2016).
14. Hsia, C. C. W., Ravikumar, P. & Ye, J. Acute lung injury complicating acute kidney injury: a model of endogenous αKlotho deficiency and distant organ dysfunction. *Bone* **100**, 100–109 (2017).
15. Mehta, R. L. et al. Sepsis as a cause and consequence of acute kidney injury: Program to Improve Care in Acute Renal Disease. *Intensive Care Med.* **37**, 241–248 (2011).
16. Coca, S. G., Singanamala, S. & Parikh, C. R. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. *Kidney Int.* **81**, 442–448 (2012).
17. Balasubramanian, G. et al. Early nephrologist involvement in hospital-acquired acute kidney injury: a pilot study. *Am. J. Kidney Dis.* **57**, 228–234 (2011).
18. Costa e Silva, V. T. et al. Nephrology referral and outcomes in critically ill acute kidney injury patients. *PLoS ONE* **8**, e70482 (2013).
19. Meersch, M. et al. Prevention of cardiac surgery-associated AKI by implementing the KDIGO guidelines in high risk patients identified by biomarkers: the PrevAKI randomized controlled trial. *Intensive Care Med.* **43**, 1551–1561 (2017).
20. Huen, S. C. & Parikh, C. R. Predicting acute kidney injury after cardiac surgery: a systematic review. *Ann. Thorac. Surg.* **93**, 337–347 (2012).
21. Chertow, G. M. et al. Preoperative renal risk stratification. *Circulation* **95**, 878–884 (1997).
22. Thakar, C. V., Arrigain, S., Worley, S., Yared, J.-P. & Paganini, E. P. A clinical score to predict acute renal failure after cardiac surgery. *J. Am. Soc. Nephrol.* **16**, 162–168 (2005).
23. Mehta, R. H. et al. Bedside tool for predicting the risk of postoperative dialysis in patients undergoing cardiac surgery. *Circulation* **114**, 2208–2216 (2006). quiz 2208.
24. Palomba, H., de Castro, I., Neto, A. L. C., Lage, S. & Yu, L. Acute kidney injury prediction following elective cardiac surgery: AKICS Score. *Kidney Int.* **72**, 624–631 (2007).

25. Aronson, S. et al. Risk index for perioperative renal dysfunction/failure: critical dependence on pulse pressure hypertension. *Circulation* **115**, 733–742 (2007).
26. Wijeysundera, D. N. et al. Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. *JAMA* **297**, 1801–1809 (2007).
27. Halford, G. S., Baker, R., McCredden, J. E. & Bain, J. D. How many variables can humans process? *Psychol. Sci.* **16**, 70–76 (2005).
28. Thottakkara, P. et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS ONE* **11**, e0155705 (2016).
29. Bihorac, A. et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann. Surg.* **269**, 652–662 (2019).
30. Koyner, J. L., Adhikari, R. & Edelson, D. P. Development of a multicenter ward–based AKI prediction model. *Clin. J. Am. Soc. Nephrol.* **11**, 1935–1943 (2016).
31. Koyner, J. L., Carey, K. A., Edelson, D. P. & Churpek, M. M. The development of a machine learning inpatient acute kidney injury prediction model. *Crit. Care Med.* **46**, 1070–1077 (2018).
32. Cheng, P., Waitman, L. R., Hu, Y. & Liu, M. Predicting inpatient acute kidney injury over different time horizons: how early and accurate? *AMIA Annu. Symp. Proc.* **2017**, 565–574 (2017).
33. Mohamadlou, H. et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can. J. Kidney Health Dis.* **5**, 1–9 (2018).
34. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
35. Meyer, A. et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir. Med.* **6**, 905–914 (2018).
36. Hosmer, D. W., Jr., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (John Wiley & Sons, 2013).
37. Praught, M. L. & Shlipak, M. G. Are small changes in serum creatinine an important risk factor? *Curr. Opin. Nephrol. Hypertens.* **14**, 265–270 (2005).
38. Brown, J. R. et al. Multivariable prediction of renal insufficiency developing after cardiac surgery. *Circulation* **116**, I139–I143 (2007).
39. Koyner, J. L., Adhikari, R., Edelson, D. P. & Churpek, M. M. Development of a multicenter ward-based AKI prediction model. *Clin. J. Am. Soc. Nephrol.* **11**, 1935–1943 (2016).
40. Mohamadlou, H. et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. *Can. J. Kidney Health Dis.* **5**, 1–9 (2018).
41. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63 (2015).
42. Naughton, C. A. Drug-induced nephrotoxicity. *Am. Fam. Physician* **78**, 743–750 (2008).
43. Mazer, M. & Perrone, J. Acetaminophen-induced nephrotoxicity: pathophysiology, clinical manifestations, and management. *J. Med. Toxicol.* **4**, 2–6 (2008).
44. Kitano, A., Motohashi, H., Takayama, A., Inui, K.-I. & Yano, Y. Valacyclovir-Induced Acute Kidney Injury in Japanese Patients Based on the PMDA Adverse Drug Reactions Reporting Database. *Drug Inf. J.* **49**, 81–85 (2014).
45. Redondo-Pachon, M. D. et al. Acute renal failure and severe thrombocytopenia associated with metamizole. *Saudi J. Kidney Dis. Transpl.* **25**, 121–125 (2014).
46. Koch, C. G. et al. Duration of red-cell storage and complications after cardiac surgery. *N. Engl. J. Med.* **358**, 1229–1239 (2008).
47. Nuis, R.-J. et al. Blood transfusion and the risk of acute kidney injury after transcatheter aortic valve implantation. *Circ. Cardiovasc. Interv.* **5**, 680–688 (2012).
48. Johnson, A. E. W. et al. Machine Learning and Decision Support in Critical Care. *Proc. IEEE Inst. Electr. Electron. Eng.* **104**, 444–466 (2016).
49. LeCun, Y. A., Bottou, L., Orr, G. B. & Müller, K.-R. in *Neural Networks: Tricks of the Trade: Second Edition* (eds. Montavon, G., Orr, G. B. & Müller, K.-R.) 9–48 (Springer Berlin Heidelberg, 2012).
50. Perez, F. & Granger, B. E. IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
51. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
52. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. McKinney, W., Others. Data structures for statistical computing in python. *Proc. 9th Python Sci. Conf.* **445**, 51–56 (2010). Austin, TX.
54. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
55. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2014)
56. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems, 2015. Software available from https://www.tensorflow.org/about/bib.
57. BRIER & W, G. Verification of Forecasts Expressed in terms of probability. *Monthey Weather Rev.* **78**, 1–3 (1950).
58. Kalton, G., Michael Brick, J. & Lê, T. Chapter VI Estimating components of design effects for use in sample design. http://citeseerx.ist.psu.edu/viewdoc/summary?doi:10.1.1.522.3221.
59. Gonzalez, E. J. & Foy, P. *Third International Mathematics and Science Study, Technical Report: Estimation of sampling variability, design effects, and effective sample sizes*. p. 87 (II, Boston College Chestnut Hill, Massachusetts, USA, 1997).
60. Kerry, S. M. & Bland, J. M. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Stat. Med.* **20**, 377–390 (2001).
61. Wolak, M. E., Fairbairn, D. J. & Paulsen, Y. R. Guidelines for estimating repeatability. *Methods Ecol. Evol.* **3**, 129–137 (2012).
62. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. In *Numerical Recipes in FORTRAN: The Art of Scientific Computing* **2**, 299–306 (Cambridge University Press, 1992).
63. Weyl, H. Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.* **77**, 313–352 (1916).
64. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
65. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 77 (2011).
66. Lele, S. R., Keim, J. L. & Solymos, P. ResourceSelection: resource selection (probability) functions for use-availability data. *R package version* 3–2 (2017). Software available at https://cran.r-project.org/src/contrib/Archive/ResourceSelection/.
67. Jiang, W. et al. Dynamic predictive scores for cardiac surgery–associated acute kidney injury. *J. Am. Heart Assoc.* **5**, e003754 (2016).

ACKNOWLEDGEMENTS

This work was supported by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A). A.M. is a participant in the BIH Charité Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. C.E. is funded by the Swiss National Science Foundation (SNSF) Ambizione Program under grant agreement no. 174025.

AUTHOR CONTRIBUTIONS

N.R. and A.M. conceived the study and drafted the manuscript. N.R., A.M., and C.E. designed the methods. N.R. and A.M. obtained the data and contributed to study design. J.K., T.K., C.S., F.S., and V.F. contributed to study design. N.R. and B.P. performed the analyses under the supervision of C.E. and A.M. All authors contributed in result interpretation and critically revised the manuscript.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

A.M. declares the receipt of consulting and lecturing fees from Medtronic GmbH and Edwards Lifesciences Services GmbH, and consulting fees from Pfizer. C.E. declares ownership of shares in codiag AG. F.S. declares the receipt of honoraria, consultancy fees or travel support from Medtronic GmbH, Biotronik SE & Co., Abbott GmbH & Co. KG, Sanofi S.A., Cardiorentis AG, Novartis Pharma GmbH. J.K. declares the receipt of lecturing fees from, Boston Scientific, LSI Solutions, Edwards, Medtronic, Abbott, Ascyrus Medical GmbH. V.F. declares (institutional) financial activities with Medtronic, Biotronik, Abbott, Boston, Edwards, Berlin Heart, Novartis, Jotec, Zurich Heart in relation to Educational Grants, honoraria, consultancy, research & study funds, fees for travel support. All other authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-020-00346-8.

Correspondence and requests for materials should be addressed to A.M.

Reprints and permission information is available at http://www.nature.com/reprints

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

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Publikationsliste

Publizierte Originalarbeiten

Rank, N., Stoiber, L., Nasser, M., Tanacli, R., Stehning, C., Knierim, J., Schoenrath, F., Pieske, B., Falk, V., Kuehne, T., Meyer, A., & Kelle, S. (2021). Assessment of 10-year left-ventricular-remodeling by CMR in patients following aortic valve replacement. *Frontiers in Cardiovascular Medicine*, *8*, 645693.
Impact Factor: 3.915

Rank, N., Pfahringer, B., Kempfert, J., Stamm, C., Kühne, T., Schoenrath, F., Falk, V., Eickhoff, C., & Meyer, A. (2020). Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *NPJ Digital Medicine*, *3*, 139.
Impact Factor: 11.653

Stoiber, L., Ghorbani, N., Kelm, M., Kuehne, T., **Rank, N.**, Lapinskas, T., Stehning, C., Pieske, B., Falk, V., Gebker, R., & Kelle, S. (2019). Validation of simple measures of aortic distensibility based on standard 4-chamber cine CMR: a new approach for clinical studies. *Clinical Research in Cardiology: Official Journal of the German Cardiac Society*, *109*(4), 454–464.
Impact Factor: 4.907

Abstracts

Rank, N., Stoiber, L., Meyer, A., Falk, V., Pieske, B., Kelle, S. (2018). Long-term left-ventricular cardiac remodeling after aortic valve replacement evaluated by advanced quantitative CMR. *Clinical Research in Cardiology, Supplement 3*.

Meyer, A., **Rank, N.**, Stoiber, L., Falk, V., & Kelle, S. (2018). Effects of arterial vs. venous revascularization of the anterior myocardial wall on long-term survival-a propensity score-based analysis. *European Heart Journal*, *39*, 126–127.

Danksagung

Hiermit möchte ich allen Personen meinen Dank aussprechen, die mich bei der Anfertigung dieser Arbeit unterstützt haben.

Zunächst möchte ich mich bei meinem Betreuer Prof. Alexander Meyer bedanken, der dieses Projekt ins Leben gerufen hat und mir durch seine positive und offene Art stets große Motivation gespendet hat.

Außerdem möchte ich Prof. Carsten Eickhoff für seine Expertise im Bereich Machine Learning, die produktiven Gespräche und seine außerordentlich hilfreichen Überarbeitungsvorschläge des Manuskripts der im Rahmen des Promotionsprojekts entstandenen Publikation danken.

Ein besonderer Dank geht an Boris Pfahringer für die fruchtbaren Diskussionen, die angenehme Arbeitsatmosphäre, seine kontinuierliche Beratung und immer hilfreiche technische Unterstützung. Ohne ihn wäre diese Arbeit nicht möglich gewesen.

Meinen Eltern und Freuden danke ich für die Ermutigungen während der Arbeit an dieser Dissertation.

Zu guter Letzt, möchte ich meinem Freund, Dr. Thomas Ewert danken, der mich auch in schwierigen Zeiten dazu motiviert hat, diese Arbeit erfolgreich fertigzustellen. Für seine Geduld und hilfreiche moralische Unterstützung bin ich ihm sehr dankbar.