Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210129

Interpretable and Continuous Prediction of Acute Kidney Injury in the Intensive Care

Iacopo VAGLIANO[a,1], Oleksandra LVOVA[a] and Martijn C. SCHUT[a]
[a] *Dept. of Medical Informatics, Amsterdam UMC, Location AMC, The Netherlands*

Abstract. Acute kidney injury (AKI) is a common and potentially life-threatening condition, which often occurs in the intensive care unit. We propose a machine learning model based on recurrent neural networks to continuously predict AKI. We internally validated its predictive performance, both in terms of discrimination and calibration, and assessed its interpretability. Our model achieved good discrimination (AUC 0.80-0.94). Such a continuous model can support clinicians to promptly recognize and treat AKI patients and may improve their outcomes.

Keywords. Acute kidney injury, clinical prediction models, machine learning, ICU

1. Introduction

Acute kidney injury (AKI) is a common and potentially life-threatening condition [1]. Clinically AKI detection uses serum creatinine increase as a marker of an acute decline in renal function. There is a lag of such an increase behind the renal injury, which results in delayed diagnosis and therefore attenuates the opportunity for early successful treatment [2]. Preventative alerts generated by medical prognosis can empower clinicians to act before a major clinical decline, improve care outcomes and optimize the use of resources [3]. As AKI occurrence in the Intensive Care Unit (ICU) is particularly high and often exceeds 50% [1], prediction of AKI in the ICU is of high relevance. Notably, to detect AKI, intensivists need to continuously monitor vital signs and laboratory measurements over time since patients' conditions may rapidly change [4].

The field of prognosis in nephrology has seen a rapid growth in machine learning applications. Machine learning might aid such continuous monitoring via a continuous prediction, meaning continuously updating the prediction of patient risk as more data become available over time. Recurrent neural networks are machine learning models particularly effective with temporal data, but they lack transparency (i.e. are often "black-boxes"), which can be a major obstacle to their practical application. In the clinical environment, models should not only make good predictions but also be interpretable [5].

To address the above-mentioned problems, we focus on the following research question: *How well can we continuously predict AKI in the ICU setting with an interpretable machine learning model?* To answer such research question, we developed a machine learning model based on Long-Short term memory (LSTM), which is a type of recurrent neural networks [6]. We internally validated its predictive performance, both in terms of discrimination and calibration, and assessed its interpretability.

[1] Corresponding Author, Dept. of Medical Informatics, Amsterdam UMC, Location AMC, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands; E-mail: i.vagliano@amsterdamumc.nl.

2. Method

2.1. Data and population

We used data from the MIMIC dataset, which contains de-identified medical information for approximately sixty thousand patient admissions to the ICUs at Beth Israel Medical Center between 2001 and 2012 [7]. We included patients with at least one measurement of serum creatinine or urine output, who were older than 18 years old, and whose length of stay in the ICU was at least 48 hours. AKI was defined according to the KDIGO guidelines [8]. We considered at most 35 days for each ICU stays (as 95% of ICU stays were shorter than 35 days): for those stays that exceeded 35 days, only records for the last 35 days were used for analysis. The final dataset consisted of 47,751 ICU stays of 34,516 unique patients.

2.2. Data preprocessing

We selected 35 variables previously indicated as potentially relevant [9, 10], excluding the ones with over 50% of missing values. The selected variables are listed in Table A1 and A2 of the supplementary material,[2] and included static (age, gender, other patients' characteristics, and administrative information) and temporal variables (physiological variables, laboratory measurements, and interventions). We did not impute missing values for the variables selected, i.e. we filled them with zeros. The missing labels were carried forward with the limit of four days to prevent using a label too old to be representative of the patient condition [10]. The remaining missing labels were filled with zeros, similarly to the input variables. After capping the extreme values at the first and 99th percentile, we normalized all numerical variables to the [0, 1] interval.

Measurements of temporal variables were made at irregular intervals and the number of measurements varied from one ICU stay to another. Therefore, we resampled such measurements at regular intervals of six hours. Each day was broken into four six-hour periods, and measurements within the same six-hour period were aggregated, using the mean for (continuous) numeric variables and the maximum value for categorical ones.

2.3. Model development

Our model for continuous AKI prediction was based on an LSTM (supplementary material, Figure A1).[2] We chose an LSTM as it is known to perform well with temporal data [11], has been successfully applied to AKI prediction [9], and allows continuous prediction. The first layer was a fully-connected embedding layer to compress the high-dimensional and sparse input variables into a lower-dimensional continuous representation, easier to manipulate by the model. The embedding layer provides the data to the three bidirectional LSTM layers, followed by another fully-connected layer to aggregate bidirectional output, and a dropout layer. The final prediction layer estimates the probability of the patient to develop AKI. The optimization function was the Binary Cross-Entropy (Bernoulli log-likelihood), and we used the Adam optimizer [12].

We compared the performance of our model with three prediction models that proved to be effective to predict AKI [13]: logistic regression, gradient boosted trees [14], and random forest [15]. The same preprocessing procedure described above

[2] https://osf.io/sjtfq, last access March 2 2021.

was applied to these models, with two differences. First, we 'flattened' our data over time by representing the values of the same input variables at every time point as different variables. Second, these models cannot provide a continuous prediction. In order to compare their predictive performance with the LSTM, the time of the prediction was 48 hours ahead of the last time point, for the ICU stays with no AKI. For stays with AKI, the time of prediction was 48 hours before the onset of AKI (i.e. before AKI occurs).

2.4. Internal validation, performance measure, and interpretability assessment

For the logistic regression, gradient boosted trees, and random forest, the data were randomly split into 90% training and 10% test sets. For the LSTM, the dataset was split into 80% training, 10% validation, and 10% test sets. For a fair comparison, we used the same test dataset used for each model. The same training set was also used but for the LSTM a validation set was retained.

We measured discrimination with the area under the receiver operating curve (AUC) and the Brier score (the mean squared error of the predictions). We assessed calibration with calibration curves. The interpretability of the LSTM was measured through the integrated gradients method [16][16].

3. Results

Table 1 outlines the discrimination (AUC and Brier score) of the LSTM and the other models. The continuous LSTM achieved the best (highest) AUC and (lowest) Brier score. Before onset, random forest showed the highest AUC, while the lowest Brier score was achieved by the gradient boosted trees. The before-onset LSTM performed worst in AUC, but it was still above 0.8. The calibration curves of the models are available in the supplementary material.[2] Random forest achieved the best calibration, the two LSTMs the worst. The feature importance of the LSTM is depicted in Figure 1. Creatinine is the major harmful risk factor and urine output is the major protective risk factor.

Table 1. AUC and Brier scores of the LSTM and the baseline models

Time of prediction	Model	AUC	Brier score
Before AKI onset	Random forest	0.93	0.128
	Gradient boosted trees	0.92	0.108
	Logistic regression	0.90	0.206
	LSTM	0.83	0.202
Continuous	**LSTM**	**0.94**	**0.101**

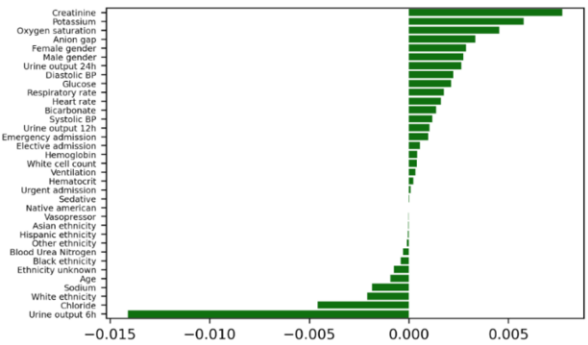

Figure 1. Features importance of the LSTM model.

4. Discussion

4.1. Main findings

The LSTM, the only continuous model, showed the best discrimination. Before the AKI onset, random forest achieved the highest predictive performance, but the LSTM yielded lower, yet competitive, results. Despite a low Brier score, the calibration curves for both the before-onset and continuous LSTMs showed room for improvement. Creatinine is the major harmful risk factor and urine output is the major protective risk factor.

4.2. Strengths and limitations

The strength of this study is proposing a continuous and interpretable model to predict AKI, applying a complete evaluation that includes discrimination, calibration, and interpretability. Our continuous model achieved an AUC of 0.94. Our study is performed on a publicly available dataset to foster reproducibility and comparison with future models. Our code is also openly available at github.com/mapo89/continuous-aki-predict.

There are some points of improvement. First, we did not include comorbidities that can be relevant for AKI. Typically, comorbidities are registered at the end of the hospital admission, so to avoid leaking information from the future, we excluded them. Second, missing values belong to the temporal variables and they likely represent measurements not performed. Such missing values are not missing at random, as in clinical practice tests are ordered based on existing observations and expectations of the clinicians. Using imputation methods such as the selection and the pattern-mixture model [17] could improve the predictive performance and reduce the risk of bias. Third, the interpretability of neural networks is not trivial. Feature importance, derived by averaging the attributes of all instances, explains which variables contributed to the predictions. However, averaging is prone to the offset of the values due to the variance on the instance level, therefore such importance is less reliable than the one derived at a model level. To better understand interpretability at the levels of layers and neurons more effort is required.

4.3. Related work

Multiple machine learning models for the AKI prediction in the ICU have been proposed. Zimmerman et al. [18] used a selection of variables from the MIMIC dataset to predict the levels of creatinine on the second and third days of ICU admission, as well as AKI thereafter. Logistic regression, random forest, and neural networks were used, with the highest AUC of 0.78. Wang et al. [19] predicted AKI in ICU with gradient boosted trees 24 hours and 48 hours before onset (AUC 0.80 and 0.77, respectively). Only two studies provide a model for the continuous prediction of AKI. Tomašev et. al. [10] focused on hospital patients and exploited a recurrent neural network. Their model provided an AUC of 0.92. Pan et al. [9] developed a continuous model for predicting AKI in the ICU, using the MIMIC dataset (AUC 0.89). Our model's discrimination was comparable to these models, while only us and Tomašev et al. assessed calibration.

However, all the above works focused on predictive performance and did not study the interpretability of the models. Freitas et al [20] discuss the validation results of an AKI prediction model for cardiac patients, based on random forest. They assessed interpretability by means of feature importance, SHAP [21], and LIME [22]. Song et al [23] developed a model to predict AKI on hospital data 48 hours before its onset and

achieved 0.81-0.87 AUC for different AKI stages. They used SHAP values for interpretability. Gong et al [24] compared several algorithms with the highest AUC of 0.77 for prediction 48 hours before AKI onset. For interpretability, they applied feature importance and SHAP. Neither of these models provides continuous AKI prediction. Notably, few continuous models have been proposed, and none studied interpretability.

5. Conclusion

This study provided a model for the continuous prediction of AKI in the ICU. Predictive performance was better than non-continuous models and interpretability was inspected. Such a continuous model can support clinicians to promptly recognize and treat deteriorating AKI patients and may improve their outcomes.

References

[1] Dennen P, Douglas IS, Anderson R Acute kidney injury in the intensive care unit: an update and primer for the intensivist. Crit Care Med, 2010;38:261-275.
[2] Aitken E, Carruthers C, Gall L, Kerr L, Geddes C, Kingsmore D. Acute kidney injury: outcomes and quality of care. QJM. 2013;106:323-32.
[3] Jonsson AJ, Kristjansdottir I, Lund SH, Palsson R, Indridason OS. Computerized algorithms compared with a nephrologist's diagnosis of acute kidney injury in the emergency department. Eur J Intern Med. 2019;60:78-82.
[4] Harty J. Prevention and management of acute kidney injury. Ulster Medical Journal 2014;83(3):149.
[5] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-1246.
[6] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 9(8):1735–1780.
[7] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi, LA, Mark RG. Mimic iii, a freely accessible critical care database. Scientific data, 2016;3:160035.
[8] Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin. Pract. 2012;120: c179–c184.
[9] Pan Z, Du H, Ngiam KY, Wang F, Shum P, Feng M. A Self-Correcting Deep Learning Approach to Predict Acute Conditions in Critical Care, CoRR, 2019.
[10] Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116-119.
[11] Paper D. Time Series Forecasting with RNNs. TensorFlow 2.x in the Colaboratory Cloud. Apress. 2020
[12] Kingma D, Ba J. Adam: A Method for Stochastic Optimization. ICLR 2014.
[13] Gameiro J, Branco T, Lopes JA. Artificial Intelligence in Acute Kidney Injury Risk Prediction. Journal of Clinical Medicine. 2020;9(3):678.
[14] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. ACM SIGKDD 2016:785–794.
[15] Liaw A, Wiener M. Classification and Regression by Random Forest. R News 2002;2(3):18–22.
[16] Mukund S, Ankur T, Qiqi Y. Axiomatic Attribution for Deep Networks. CoRR, 2017.
[17] Glynn RJ., Laird NM, Rubin DB. Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse. In Drawing Inferences from Self-Selected Samples, 1986;115–42.
[18] Zimmerman LP, Reyfman PA., et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. BMC Med Inform Decis Mak 2019;19:16.
[19] Wang Y, Wei Y, Wu Q, Yang H, Li J. An Acute Kidney Injury Prediction Model Based on Ensemble Learning Algorithm, Int. Conf. on Information Technology in Medicine and Education, 2019;18-22.
[20] Freitas da Cruz H et al. Using Interpretability Approaches to Update "Black-Box" Clinical Prediction Models: an External Validation Study in Nephrology, Artificial Intelligence in Medicine, 2021;111.
[21] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. in Advances in neural information processing systems, 2017.
[22] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;1135–1144.
[23] Song X, Yu ASL, Kellum, JA et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 2020;11:5668.
[24] Gong K, Lee HK, Yu K, Xie X, Li J. A prediction and interpretation framework of acute kidney injury in critical care, Journal of Biomedical Informatics, 2021:113.