Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220683

External Validation and Transportability of Models to Predict Acute Kidney Injury in the Intensive Care Unit

Iacopo VAGLIANO[a,1], Carmen BYRNE SALSAS[a], Tina WÜNN[a] and Martijn C. SCHUT[a]
[a] *Dept. of Medical Informatics, Amsterdam UMC, Location AMC, The Netherlands*

Abstract. External validation of models for the prediction of acute kidney injury (AKI) is rare. We externally validate AKI prediction models in intensive care units. The models were developed on the Medical Information Mart for Intensive Care dataset and validated on the eICU dataset. Traditional machine learning models show limited transportability to the new population (AUROC < 0.8). Models based on recurrent neural networks, which can capture complex relationships between the data, transport well to the new population (AUROC 0.8-0.9). Such models can help clinicians to recognize AKI and improve the outcome.

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

1. Introduction

Many prediction models are developed but are typically only internally validated. This gives insight into the model's performance on new patients from the same target population. However, external validation is needed to assess how well the model performs on new patients from a different population than the one used to develop the model, e.g., from a different hospital. Debray et al. proposed a new framework for the interpretation of external validation studies of clinical prediction models [1].

Such lack of external validation holds also for models that predict acute kidney injury (AKI), hampering their implementation in clinical practice. AKI is common in intensive care unit (ICU) patients and is defined as an abrupt decrease in kidney function characterized by a sudden increase in serum creatinine or a reduction in urine volume [2]. Clinicians rely on serum creatinine increase to mark an acute decline in renal function and detect AKI, but the diagnosis is delayed because there is a lag of such an increase behind the renal injury. This lag lessens 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]

Our aim is to externally validate five machine learning models to predict AKI in ICU patients. We have previously developed models [4] on the Medical Information

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

Mart for Intensive Care (MIMIC) dataset [5] and we externally validate their performance on patients from the eICU dataset [6].

2. Method

2.1. Data and population

The multi-center eICU Collaborative Research Database contained clinical data for over 200,000 ICU admissions and 139,367 unique patients across the United States [6]. Patients and variables were included for analysis and processed as done for model development [4]. Data and preprocessing are explained in the supplementary material.[2]

2.2. Prediction Models

We externally validated models developed on the MIMIC data in our previous work [4]. Such models were logistic regression, gradient boosted trees [7], random forest [8], and two variants of a Long-Short term memory (LSTM), a type of recurrent neural networks [9]. One variant enables a continuous prediction, Which means continuously updating the prediction of patient risk as more data become available over time. The other LSTM variant and all other models predict AKI before onset, i.e. before AKI occurs. The time of the prediction for these models was 48 hours ahead of the last time point, for the ICU stays with no AKI, or 48 hours before the onset of AKI for stays with AKI.

2.3. External validation and performance measures

Debray et al. described a framework for the interpretation of external validation studies of clinical prediction models [1]. If the populations have different case mix, the model transportability is assessed. To check how similar the development and validation populations were, we developed a membership model based on logistic regression, which predicted the probability that an individual belongs to the development population (MIMIC) or the validation population (eICU). If the membership model performs well, the model transportability is assessed. The membership model used the same variables as the validated models plus the AKI outcome. Its discrimination was assessed by the average area under the receiver operating curve (AUROC) in a 10-fold cross-validation.

To assess the performance of the validated models, we used the same measures as in the model development [4]. We measured discrimination with the AUROC, the Brier score and the area under the precision-recall curve (AUPRC). AUPRC was added because of class imbalance, being more informative on imbalanced data [10]. Calibration was assessed with calibration curves.

3. Results

The dataset consisted of 142,432 unique ICU stays. Descriptive statistics of the population are in Section 2 of the supplementary material.[2]

[2] https://osf.io/9gy8z/, last access April 20, 2022.

The AUROC of the membership model was 0.93 with a standard deviation of 0.002, indicating that transportability was assessed. Table 1 outlines the discrimination of the five prediction models. Logistic regression and random forest performed poorly, gradient boosted trees fairly, the before-onset LSTM well, showing the best AUPRC, and the continuous LSTM excellent, achieving the best AUROC and Brier score.

Table 1. Discrimination of the externally validated models in terms of AUROC, AUPC and Brier score

Time of prediction	Model	AUROC	AUPRC	Brier score
Before AKI onset	Logistic regression	0.62	0.40	0.308
	Random forest	0.62	0.39	0.259
	Gradient boosted trees	0.78	0.68	0.208
	LSTM	0.82	**0.81**	0.124
Continuous	LSTM	**0.93**	0.19	**0.075**

Figure 1 shows the calibration curves of the models on eICU data. Overall, the models are not well calibrated. Random forest shows the best calibration. However the LSTM models, notably the before-onset one, show better calibration for most of the predictions (the before-onset LSTM is even better than random forest for most of the predictions).

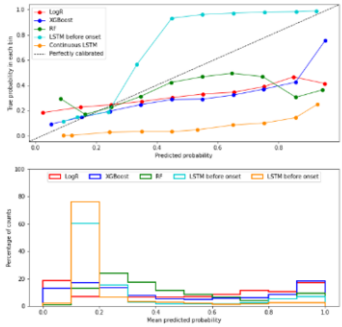

Figure 1. Calibration curves of the validated models. The histograms are normalized to the number of predictions of each model. LogR is logistic regression, RF random forest, XGBoost gradient boosted trees.

4. Discussion

In this study, we assessed case mix similarity to establish models' transportability between the MIMIC and eICU populations. We externally validated five models on eICU. The membership model's results (AUROC 0.93) indicated that the two populations have different case mix. Thus, the external validation assessed the models' transportability.

The LSTM models showed good to excellent discriminative performance in eICU (AUROC 0.82-0.93), which suggested good model transportability. The other models achieved poor to fair performance (AUROC 0.60-0.8). The lower results of these models might stem from differences in the eICU data compared to the MIMIC data. Notably, MIMIC variables had less than 50% of missing data. Most of these variables in eICU have more missing values (see Table S2 in the supplementary material[2]). LSTM models can capture more complex relationships between the data and the outcome, which may have helped them to better transport to a new population.

Calibration assessment showed room for improvement. LSTMs showed better calibration for most of the prediction. While recalibrating the models is beyond the scope of this paper, a recent study proposed neural networks' recalibration without the need of retraining the models [11] and may constitute future work.

The strength of this study is using the framework proposed by Debray et al. to analyze and interpret the results from this external validation study [1]. We used two publicly available datasets to promote reproducibility of our study. Our code is available at bitbucket.org/aumc-kik/aki-models-external-validation. There are some limitations to this study. First, eICU has generally more missing values than MIMIC. Second there are some differences in variable representation, e.g. the admission type was missing in eICU. This loss of information in the eICU data could have led to worse predictions.

Various models for the AKI prediction in the ICU have been proposed, but external validation is rare. Da Cruz et al. developed models on MIMIC and validated them on a cohort of the Mount Sinai Health System in New York [12]. The AUROC dropped from 0.81 to 0.64 for their logistic regression, and from 0.88 to 0.73 for their random forest. Meyer et al. validated models developed on a German tertiary care center for cardiovascular diseases with MIMIC patients (AUROC of 0.91 with the best model, which was a recurrent neural network) [13]. Schneider et al. developed a decision tree with an accuracy of 80% for the development and 73% for the validation population [14].

5. Conclusions

External validation is crucial to assess how models perform on a new population, but is uncommon. We validated five AKI prediction models. LSTM can capture complex relationships between the data and the outcome and transport well to the new population. Such models can aid clinicians to promptly recognize AKI and improve the outcome.

Acknowledgments: We thank Evani Lachmansingh for the eICU data extraction.

References

[1] Debray TP, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015 Mar;68(3):279-89.
[2] Aitken E, Carruthers C, Gall L, Kerr L, Geddes C, Kingsmore D. Acute kidney injury: outcomes and quality of care. QJM. 2013 Apr;106(4):323-32.
[3] Jonsson AJ, et al. Computerized algorithms compared with a nephrologist's diagnosis of acute kidney injury in the emergency department. Eur J Intern Med. 2019 Feb;60:78-82.
[4] Vagliano I, Lvova O, Schut MC. Interpretable and Continuous Prediction of Acute Kidney Injury in the Intensive Care. Stud Health Technol Inform. 2021 May 27;281:103-107.
[5] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.
[6] Pollard TJ, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018 Sep 11;5:180178.
[7] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).
[8] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002 Dec 3;2(3):18-22.
[9] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-80.
[10] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015 Mar 4;10(3):e0118432.
[11] Toubeau JF, et al. Recalibration of recurrent neural networks for short-term wind power forecasting. Electric Power Systems Research. 2021 Jan 1;190:106639.
[12] Da Cruz HF, et al. Using interpretability approaches to update “black-box” clinical prediction models: an external validation study in nephrology. Artificial Intelligence in Medicine. 2021 Jan 1;111:101982.
[13] Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018 Dec;6(12):905-914.
[14] Schneider DF, Dobrowolsky A, Shakir IA, et al. Predicting acute kidney injury among burn patients in the 21st century: a classification and regression tree analysis. J Burn Care Res. 2012;33(2):242-51.