Nephrol Dial Transplant (2020) 35: 204–205
doi: 10.1093/ndt/gfz226
Advance Access publication 28 November 2019

EDITORIAL

Prediction of acute kidney injury using artificial intelligence: are we there yet?

Wim Van Biesen [1,2], Jill Vanmassenhove[1] and Johan Decruyenaere[2,3]

[1]Renal Division, Ghent University Hospital, Ghent, Belgium, [2]Justifiable Digital Health Consortium, Ghent University Hospital, Ghent, Belgium and [3]Department of Intensive Care, Ghent University Hospital, Ghent, Belgium

Correspondence to: Wim Van Biesen; E-mail: Wim.vanbiesen@ugent.be

Management of acute kidney injury (AKI) is suboptimal and often opportunities for AKI prevention are missed [1]. AKI is frequently associated with other underlying conditions and will, in most cases, be handled by non-nephrologists in the hospital setting who are less experienced in diagnosing AKI. This has made the development of detection or prediction models of AKI a hot topic. Most of these models are restricted to estimating the risk for AKI at baseline, e.g. on admission or preoperatively, and/or in specific settings, e.g. cardiac surgery or sepsis, thus hampering their generalizability [2]. Recently a machine learning-based algorithm was shown to perform as well as physicians in predicting AKI stages 2–3 on the day of admission, albeit in the restricted setting of the intensive care unit (ICU), on the basis of which the algorithm was developed [3]. More advanced prediction models have shown increased clinical usefulness by allowing continuous, and nearly real-time, risk prediction, taking into account longitudinal patient data, and thereby the dynamic status of the patient, thus making implementation of an (automated) electronic alert system highly attractive [4]. In addition, many models predict only advanced stages of AKI or the need for renal replacement therapy, thus missing the opportunity to prevent or mitigate AKI.

In *Nature*, Tomašev *et al.* [5] reported the development of a continuous prediction model for AKI based on artificial intelligence (AI). The authors used longitudinal data from electronic health records of >700 000 inpatients, as well as outpatients, across all specialties to train a deep learning recurrent neural network model. The system was trained using not only current medical data, but also previous data for up to 2 years before admission, resulting in a flabbergasting 6 billion independent data entries. For every case, presence or absence of AKI was labelled to allow supervised learning. The resulting model was able to predict AKI in 55.8% of all inpatient cases of diagnosed AKI, with a lead time of up to 48 hours and a ratio of two false alerts for every true alert.

This project illustrates the potential of using AI trained in big data in medicine. However, it also reveals the limitations and pitfalls of such an approach and the issues that need further research and attention.

First, the performance of the model is not that impressive. The model has a low sensitivity of 55.8%, meaning half of the AKI episodes are missed. This may be a deliberate choice to inflate the specificity. In settings with a low prevalence of AKI, even highly performing models will have to strike a balance between high sensitivity, and thus missing fewer cases, and specificity, and thus reducing false positives and tackling alert fatigue. Further research is needed to explore which approach will, in clinical practice, result in substantial improvement in outcomes [6]. Of particular note, in a general hospital setting, missing AKI cases might be considered more problematic than alert fatigue. However, it is important to consider that low sensitivity might simply be due to the fact that a large proportion of AKI cases simply *cannot* be predicted, just like one cannot predict the side on which a coin will land in a coin toss experiment. These drawbacks are inherent to the laws of probability and hence apply to all diagnostic and prediction models, whether traditional or machine learning based [7].

Tomašev *et al.* [5] used supervised learning, in which input features are associated with pre-specified output labels based on a mathematical algorithm. This implies that during training the model is provided with the correct label for each case and also during training the model learns from its 'mistakes' by adapting the weights that associate the data with the label in the algorithm. Such an approach presumes that the categorization used is: (i) 'transparent' and 'uniform', implying everybody understands clearly and unequivocally what the category label represents exactly; (ii) 'relevant', with categories making a meaningful distinction between cases; (iii) 'unique', meaning that every case belongs to one, and only one, category and (iv) 'exhaustive', meaning that all cases that exhibit meaningful differences can be assigned to a different category. Last, it also presumes that all cases in the training set are labelled correctly. When these assumptions are violated, the resulting algorithm will achieve a poor performance in practice.

Most AKI prediction algorithms claim to be based on Kidney Disease: Improving Global Outcomes criteria, but in reality they are based only on the creatinine level, and not on the urinary output criterion. This implies AKI predicted by these models might differ from AKI as understood in clinical practice and, as such, the definition of what is predicted by the algorithm is not transparent nor uniquely defined. Importantly, urinary output could represent the cheapest continuously available predictor of AKI we have, and instructing healthcare staff to monitor diuresis might be more effective in improving AKI management than searching for new AKI biomarkers or developing prediction algorithms for AKI [8].

An electronic alert for AKI would have the most added value in those settings with the highest risk for missing an AKI diagnosis. However, in a retrospective data set, cases where a diagnosis of AKI was not considered by the physician are highly likely to lack the data to ground the diagnosis. One can opt to consistently label these cases as 'no AKI', as in Tomašev *et al.*'s paper [5]. In this case, the label of 'no AKI' would have no relevance since it does not distinguish between 'confirmed no AKI', as grounded in the available data, and 'uncertain AKI', i.e. data are not available to ground the diagnosis. 'Uncertain AKI' might represent missed AKI cases or incorrect labelling in the training set, thus resulting in false negatives in the test set through supervised learning—in this way, cases missed by the physician will also be missed by the prediction model. One could also opt to exclude cases that miss the data to ground the diagnosis. However, this would mean the training set will not include cases of 'easily missed AKI' and, as a consequence, will not be exhaustive. The typical pattern of these cases will not be recognized in the test set, resulting in their random classification as either AKI or no AKI.

Furthermore, the model is not only trained in using test results, but also in the use of metadata, such as whether tests are ordered, as well as their timing, regardless of the test results. There is evidence these metadata are more informative than actual test results for predicting outcomes, including survival [9]. However, such metadata rely heavily on the expertise of the physician who will order the test *because* of a presumed risk of AKI. This will result in the creation of a vicious circle within the model. During training the model will blindly (i.e. without any knowledge about the problem) associate the request of the physician for a test to detect AKI with the later occurrence of AKI. The model will thus generate an alert for a potential AKI problem that the physician already recognized himself.

Furthermore, implementing such a model could have other unexpected 'side effects' in daily clinical practice. If a user strongly trusts in the performance of the electronic alert system, they might choose to 'wait' for an alert trigger from the model before taking action, while the model, in fact, 'needs' these actions to be taken first so it can estimate the risk of AKI and trigger the alert. This may, of course, have detrimental effects on timely AKI recognition. Alternatively, users can order, as a default setting in practice, all the necessary tests to feed the algorithm for all patients. However, in this default setting, the relationship between the results of these tests and AKI will shift due to a change in the pre-test probability distributions between the training set and the clinical practice data set. The model might not be calibrated to this new setting. In addition, it can also lead to the well-known 'neural net tank urban legend' [10], as illustrated in an AI-based chest X-ray diagnosis model [11] in which the diagnosis was largely driven by the type of X-ray machine used, with differences in the machine used in the outpatient setting (low incidence of pneumonia) compared with the ICU setting (high incidence of pneumonia).

Last, an alert trigger needs to be timely to avoid or mitigate the risk for AKI. Although a 48-hour prediction window would seem adequate, not all AKI diagnoses are predicted within 48 hours before the onset of AKI by the prediction system. In fact, only 20% of AKI cases are predicted >24 hours before AKI onset [5]. For the majority of cases, this might prove to be too late for an effective intervention [2]. In addition, in the case of AKI, intervention would consist of preventive measures that should be provided to all patients, not only to those with impending AKI.

The points mentioned above could unveal the Achilles tendon of this type of prediction models based on retrospective observational datasets, and could jeopardize their clinical usefulness.

FUNDING

W.V.B. is supported by a research grant from Fonds Wetenschappelijk Onderzoek (FWO.OPR.2019.0045.01).

CONFLICT OF INTEREST STATEMENT

None declared. The content of this article has not been published previously in whole or part, also not in abstract format.

REFERENCES

1. Stewart JA. Adding insult to injury: care of patients with acute kidney injury. *Br J Hosp Med (Lond)* 2009; 70: 372–373
2. Vanmassenhove J, Kielstein J, Jorres A *et al.* Management of patients at risk of acute kidney injury. *Lancet* 2017; 389: 2139–2151
3. Flechet M, Falini S, Bonetti C *et al.* Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. *Crit Care* 2019; 23: 282
4. Colpaert K, Hoste EA, Steurbaut K *et al.* Impact of real-time electronic alerting of acute kidney injury on therapeutic intervention and progression of RIFLE class. *Crit Care Med* 2012; 40: 1164–1170
5. Tomašev N, Glorot, X, Rae JW *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; 572: 116–119
6. Kane-Gill SL, O'Connor MF, Rothschild JM *et al.* Technologic distractions (part 1): summary of approaches to manage alert quantity with intent to reduce alert fatigue and suggestions for alert fatigue metrics. *Crit Care Med* 2017; 45: 1481–1488
7. Shah NH, Milstein A, Bagley, SC. Making machine learning models clinically useful. *JAMA* 2019; 322: 1351–1352
8. Kolhe NV, Reilly T, Leung J *et al.* A simple care bundle for use in acute kidney injury: a propensity score matched cohort study. *Nephrol Dial Transplant* 2016; 31: 1846–1854
9. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018; 361: k1479
10. *The Neural Net Tank Urban Legend.* https://www.gwern.net/Tanks (25 September 2019, date last accessed)
11. Couzin-Frankel J. Medicine contends with how to use artificial intelligence. *Science* 2019; 364: 1119–1120

Received: 20.8.2019; Editorial decision: 14.10.2019