Machine Learning to Predict Acute Kidney Injury

F. Perry Wilson

Program of Applied Translational Research, Yale School of Medicine, New Haven, CT.

The widescale adoption of electronic health record (EHR) technology has led to an unprecedented accumulation of medical data, such that petabytes of patient information are now easily accessible to computer systems. Data have inherent value, as evidenced by the astounding success of technology companies that rely primarily on the exchange of data to generate profit.[1] However, data science in health care has been stunted compared with other industries. This is due in no small part to limitations in accessing health data due to concerns regarding privacy, questions over data ownership, and uncertainties around applicability.

However, increasing interest from industry may be the key to leveraging patient-level data in ways that can improve care and advance science. A recent publication by Tomasev et al[2] in Nature shows how a collaboration between experienced machine-learning researchers (in the form of DeepMind, a Google subsidiary) and clinicians can create impressive results. The objective of the study was rather straightforward: to develop a computational model that, fed data inputs from a given individual, would output their likelihood of developing acute kidney injury (AKI) in the near future. Studies have done this before, but none with so vast a data set or such accurate results.[3,4]

What Does This Important Study Show?

In terms of the sheer number of data points analyzed, this is one of largest studies of machine learning in medicine to date. Researchers leveraged the US Veterans Affairs (VA) clinical database, creating a data set of more than 700,000 individuals across 1,239 health care facilities. Each patient contributed multiple time-stamped datapoints—a total of more than 6 billion clinical-event entries.

At each timepoint during an inpatient stay, entries were classified as being within 24, 48, or 72 hours before AKI. Instead of asking the question "Will this patient ever develop AKI in the future?" the researchers asked "Will this patient develop AKI in the near future?" The latter question is much more clinically relevant (because an AKI early-warning system could potentially lead to preventative action), though the choice complicates modeling substantially.

Using a recurrent neural network, a machine-learning algorithm that has some advantages with regard to longitudinal data, the researchers created a time-updated prognostic model.[5]

The area under the receiver operator characteristic curve, a standard metric for assessing the performance of prognostic models, was 0.92 for prediction of AKI within the next 48 hours. This performance indicates that given a patient who would develop AKI within 48 hours and one who would not, the model would give the patient who would develop AKI the higher score 92% of the time. Other studies, even those that leverage novel serum and urine biomarkers to predict AKI, have rarely exceeded 75% to 80% at this prediction task.[6]

However, in the real world, implementation of these models rests on determining a model cut point, or threshold at which some action could be taken. High-specificity cut points (enriching for patients very likely to develop AKI) may be appropriate for interventions that carry some risk (such as empirical volume resuscitation), while high-sensitivity cut points (ensuring that few patients with impending AKI are missed) might be appropriate for more benign interventions (such as serum creatinine [Scr] monitoring). The authors specify a cut point that would capture 55.8% of all AKI cases within a 48-hour window. At that threshold, there would be 2 false predictions for every 1 positive prediction and just <3% of hospitalized patients would alert on a daily basis, making it appropriate for certain low-cost but high-yield interventions (the authors suggest a "clinical assessment").

Higher cut points would trigger less often and lead to fewer false-positive results but miss more patients who go on to develop AKI. These tradeoffs are inherent to prognostic modeling, and practical decisions about when to alert rest critically on the specific interventions the alerts are supposed to trigger. A less complex algorithm with poorer prognostic performance, it should be noted, would exacerbate this issue by firing more often and would lead to more false-positive results for the same AKI capture rate.

The impressive complexity of the algorithm is a double-edged sword. Although machine-learning approaches with extensive data preprocessing and the integration of thousands of clinical variables can show impressive results in retrospective data sets, prospective implementation presents novel challenges. The study used 620,000 features (variables) as inputs to the model. Any prospective study would need to capture those hundreds of thousands of inputs to output an appropriate prediction at a given time point. In real time, any one of these inputs can "break," leading to degradation or even failure of the predictive algorithm. As a simple example, a recent upgrade of our laboratory system led to a renaming of the Scr field in our EHR. Because this term is used for a variety of studies, including those of AKI alerts,[7] until the new name was updated in the computer code, no Scrbased alerts fired. In this case, the cause was quickly identified and the fix was straightforward. With thousands of input variables, constant curation will be necessary to ensure that unrelated changes in data structures do not break the predictive model.

On an even more basic level, not all predictors are available in all data sets. While common features (such as serum potassium level) are well represented in diverse health systems, there are likely hundreds of features that can be captured in the VA, but not in other institutions. This is particularly true of outpatient data, a critical component of the predictive model that the authors present and one that may be the key to explaining the increased predictive power seen in this model versus prior efforts. It is worth noting that the study only included patients

with at least 1 year of outpatient data, limiting generalizability to health systems with less complete data capture than the VA.

Real-time application of these models also requires real-time computing power, and many institutions will want to perform these calculations on site to avoid sending potentially sensitive information to third parties. The impact on the day-to-day functioning of information technology systems (often strained under ordinary circumstances) is difficult to quantify.

A solution to all these issues involves feature selection: machine-learning approaches that create models that maximize predictive power with a sparse set of inputs. Assume that the team charged with EHR implementation of an AKI prognostic model believes that a maximum of 20 time-updated inputs can be coded and monitored in perpetuity. Given 600,000 potential inputs, there are 1.5×10^{97} models that could be developed with 20 inputs. Obviously, a brute-force approach to evaluation of these models is infeasible in any conceivable timescale (barring breakthroughs in quantum computing), but powerful feature-selection algorithms, such as evolutionary search functions and population-based incremental learning, can explore the space of possible models efficiently and deliver results on a timescale approaching days, instead of eons.[8,9]

Finally, we are not given details on the contribution of time-variant versus time-invariant features to the overall prediction. In other words, are individuals alerting because they are at high risk from the moment they enter the hospital (based on an elevated baseline Scr, the presence of diabetes, or other chronic conditions) or because of the dynamic changes that take place during hospitalization (an increasing potassium level, a decreasing bicarbonate level)? If the former, implementation will lead to "admission" alerts, which may be disregarded by providers amidst the slew of other tasks required to admit a patient to the hospital.

How Does This Study Compare With Prior Studies?

Prognostic modeling studies are often compared by the performance of the prognostic model, and by that metric, this study certainly outpaces those that have come before.[3,4,10] This is likely due more to the data availability than the particular algorithm used. Although the area under the curve for the primary model (a recurrent neural network) was 0.92, the authors report that a simple logistic regression with the same inputs yielded an area under the curve of 0.86, which still outclasses prior work. This indicates an emerging truism of prognostic modeling: the data matter more than the algorithm.

Rather than model performance, the main contribution of the study may surround identification of appropriate hyperparameters for machine learning in this space. In the context of machine learning, a "hyperparameter" represents a discrete value or choice made during the modeling process. For example, in creating a neural network, a data scientist may choose how many layers there should be, how many neurons in each layer, and how the neurons should connect. The machine-learning algorithm itself can be considered a hyperparameter: should we use a recurrent or convolutional neural network? A logistic

regression or a decision tree? Permuting these decisions "tunes" the model, but as these choices multiply, the space of possible tunings explodes and it rapidly becomes challenging to evaluate them all. The resources of DeepMind allowed evaluation of a vast number of hyperparameters, many more than could be considered by a typical university data science laboratory. These parameters can be used to inform the starting point for future efforts, bypassing the laborious process of hyperparameter selection. In effect, DeepMind has given researchers a standard tuning.

What Are the Implications for Nephrologists?

While prospective implementation of a complex model like this remains challenging, should it become feasible, the obvious question is "what do we do now?" The history of AKI clinical research is littered with failed therapies targeted to individuals after AKI has been diagnosed through Scr level.[11,12] Although there is reason to hope that targeting therapies earlier, before Scr level even increases, may improve outcomes, we have few empirical data to prove this is the case. Nevertheless, the adoption of real-time prognostic models like this will form the basis of these research efforts in the future.

Nephrologists in the near future may receive "pre-AKI" consults, which present a unique clinical opportunity. The management of these patients will at first follow the principles we use in AKI consults now; management of hemodynamics and avoidance of nephrotoxins seem to be appropriate responses. Of course, the risk-benefit ratio of certain interventions is fundamentally changed when the patient does not yet have the condition for which we are consulted. Stopping treatment with an aminoglycoside may be reasonable in the throes of AKI, but can the same be said when kidney function is normal, particularly if the drug has some well-described benefit?

There is potential for overreaction as well. If a patient is ready for discharge but the pre-AKI alert fires, should they continue to be monitored? Given the frequency and usually self-limited nature of AKI, a less-is-more approach should rule the day until such time as high-quality data inform the management of pre-AKI patients. Trials randomizing pre-AKI alerting to usual care would seem particularly valuable at this point.

Due to the computationally simple definition of AKI (diagnosable with one easily measured clinical parameter), prognostic modeling is feasible and perhaps inevitable. As machine-learning approaches become more sophisticated, parsimonious high-performing models will enter our daily practice. When they do, nephrologists must be prepared to answer the call, even if the telephone has not yet rung.

Support:

References

1. Parkins D The world's most valuable resource is no longer oil, but data. Economist. 2017.

2. Tomasev 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. [PubMed: 31367026]

3. Simonov M, Ugwuowo U, Moreira E, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: a descriptive modeling study. PLoS Med. 2019;16(7):e1002861. [PubMed: 31306408]

4. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. Crit Care Med. 2018;46(7):1070–1077. [PubMed: 29596073]

5. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2016;24(2):361–370.

6. Ho J, Tangri N, Komenda P, et al. Urinary, plasma, and serum biomarkers' utility for predicting acute kidney injury associated with cardiac surgery in adults: a meta-analysis. Am J Kidney Dis. 2015;66(6):993–1005. [PubMed: 26253993]

7. Mutter M, Martin M, Yamamoto Y, et al. Electronic Alerts for Acute Kidney Injury Amelioration (ELAIA-1): a completely electronic, multicentre, randomised controlled trial: design and rationale. BMJ Open. 2019;9(5):e025117.

8. Vafaie H, De Jong K, eds. Genetic algorithms as a tool for feature selection in machine learning Presented at: Proceedings Fourth International Conference on Tools with Artificial Intelligence; 11 12, 1992; Arlington, VA.

9. Baluja S Population-Based Incremental Learning A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning. Pittsburgh, PA: Department of Computer Science, Carnegie-Mellon University; 1994.

10. Mohamadlou H, Lynn-Palevsky A, Barton C, et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. Can J Kidney Health Dis. 2018;5 2054358118776326. [PubMed: 30094049]

11. Bellomo R, Haase M, Ghali W, Bagshaw S, Delaney A. Loop diuretics in the management of acute renal failure: a systematic review and meta-analysis. Crit Care Resusc. 2007;9(1):60–68. [PubMed: 17352669]

12. Kellum JA, Decker JM. Use of dopamine in acute renal failure: a meta-analysis. Crit Care Med. 2001;29(8):1526–1531. [PubMed: 11505120]