0196-0644/$-see front matter

Selection Bias in the Predictive Analytics With Machine-Learning Algorithm

To the Editor:

I read with great interest the study by Martinez et al,[1] who developed a machine-learning model for the early prediction of acute kidney injury. The model was created in the emergency department (ED) and can predict acute kidney injury development at 24, 48, and 72 hours after the index ED visit. The model showed good performance in terms of discrimination. However, model calibration was not performed, leaving unknown whether the model predicts equally well at various risk strata.

First, one important issue in this study was that the patient selection was problematic. The authors enrolled individuals with a serum creatinine (sCr) measurement in ED and subsequent measurement of sCr at specific windows, and those with an initial level of greater than 4 mg/dL were excluded. It is reasonable to deduce that patients without sCr measurement during the ED stay are those considered to have normal renal function as judged by the treating physician, which can cause selection bias. The health care process of laboratory measurement has been documented in other literature.[2] The missing value contains prognostic information.[3] For instance, hospitalized patients without blood gas measurement are more likely to have better prognosis than those with such a measurement. A better approach is to impute baseline sCr level by the 4-variable Modification of Diet in Renal Disease equation.[4,5]

Second, laboratory measurements obtained at the initial encounter at ED admission are usually those at the worst stage, and the condition typically improves after appropriate treatment. Thus, the baseline sCr level obtained at the initial ED visit is not really the "baseline," but rather the maximum value (eg, sCr level returns normal after appropriate volume expansion). The baseline sCr for individuals without preexisting renal dysfunction should be better predicted with the Modification of Diet in Renal Disease equation. The authors excluded patients at sCr level greater than 4 mg/dL, which is a very high value, and patients with existing acute kidney injury at the ED visit were excluded. For example, a patient can have acute kidney injury at the ED visit with an sCr level of 2.1 mg/dL, and renal function can gradually be improved by effective treatment. This patient would not be considered as having acute kidney injury in the prediction model. More strictly, such patients should be considered as those without renal function deterioration. The model aims to predict acute kidney injury at an early stage, indicating the target population should have no acute kidney injury at the ED visit. Thus, I suggest excluding patients with elevated sCr level at the ED visit.

Third, a final problem with the study was the use of metrics for evaluating model performance for imbalanced outcome data. The incidence of acute kidney injury in the study population was less than 10%; the balance between missed cases (false-negative results) and misdiagnosed cases (false-positive results) should be considered according to the clinical problem. The recall-precision curve and relevance area under receiver operating characteristic curve should be reported.

Jiyuan Jiang, MD
Department of Emergency Medicine
Affiliated Dongyang Hospital of Wenzhou Medical University
Dongyang, Zhejiang, People's Republic of China

https://doi.org/10.1016/j.annemergmed.2020.09.004

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The author has stated that no such relationships exist.

1. Martinez DA, Levin SR, Klein EY, et al. Early prediction of acute kidney injury in the emergency department with machine-learning methods applied to electronic health record data. *Ann Emerg Med*. 2020;76:501-514.
2. 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.
3. Zhang Z, Zhao Y, Canes A, et al; on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med*. 2019;7; 152.
4. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med*. 2009;150:604-612.
5. Zhang Z. Missing data imputation: focusing on single imputation. *Ann Transl Med*. 2016;4:9.

In Reply:

We appreciate Dr. Jiang's interest in our research and the thoughtful comments related to patient selection, model calibration, and performance evaluation metrics. We described the development and evaluation of machine-learning models to predict short-term risk for acute kidney injury in the emergency department (ED) setting.[1] This study serves as a proof of concept that electronic health record data, collected very early in the ED encounter, can be used to generate reliable estimations of short-term risk for kidney injury. Our ultimate objective is to incorporate model-based predictions into a clinical decision support (CDS) system that enables better management and outcomes for patients with or at risk for acute kidney injury. Within the context of a complete CDS system, predictive data will be used in combination with surveillance and detection algorithms to target patients with or at risk for acute kidney injury with renally protective CDS.

In regard to patient selection, we concede that all models derived from electronic health record data are subject to some element of selection bias. However, we believe the major limitation of our model is selection based on presence of outcome data (eg, repeated serum creatinine [sCr] measurement), as discussed in our original "Limitations" section.[1] Selection based on sCr measurement during the ED encounter is less problematic because this occurred for greater than 98% of patients hospitalized from our study sites. As pointed out by Dr. Jiang, our use of sCr measured on ED arrival to establish baseline introduces potential for inclusion of patients with community-acquired acute kidney injury and could lead to underestimation of overall acute kidney injury incidence. Although imperfect, this approach is commonly used for estimation of hospital-acquired acute kidney injury incidence and development of hospital-based acute kidney injury prediction models.[2,3] Alternative approaches exist, including imputation of baseline using the Modification of Diet in Renal Disease equation as suggested. This is also problematic because it may overestimate acute kidney injury in patients with chronic kidney disease. Exclusion of all patients with elevated sCr levels, also suggested, would preclude identification of acute kidney injury in those with chronic disease, a particularly at-risk population. We posit that these challenges are best addressed through the development of CDS platforms that leverage prediction and detection algorithms simultaneously, as we intend to do.

We also acknowledge the concern raised about the study's emphasizing discriminative performance over calibration. Although maximization of calibration can lead to degradation of discrimination,[4] we agree these must be balanced and do always consider calibration before model deployment. Our focus on discriminative performance is driven by our intention to generate clear treatment recommendations when risk of acute kidney injury exceeds a threshold, rather than display probabilistic estimates directly to clinician-users. The latter are more affected by calibration error. Precision-recall curves, as suggested by Dr. Jiang, may be the optimal approach for relating the model to clinical utility, especially given the class imbalance (ie, low acute kidney injury incidence) in our data set. This is exactly what was done. Multiple operating points and associated precision-recall curves for predicting stage 1 acute kidney injury were reported in the original submission's supplementary data.[1] Although sensitivity and specificity declined as the prediction horizon widened, classifiers' precision increased in parallel with acute kidney injury's cumulative incidence.

Diego A. Martinez, PhD
Scott R. Levin, PhD
Jeremiah S. Hinson, MD, PhD
Department of Emergency Medicine
Johns Hopkins University
Baltimore, MD

https://doi.org/10.1016/j.annemergmed.2020.09.005