Published in final edited form as:
Crit Care Med. 2018 July ; 46(7): 1196–1197. doi:10.1097/CCM.0000000000003170.

Towards the ‘Plateau of Productivity’: enhancing the value of machine learning in critical care

Vincent X Liu, MD MS
Kaiser Permanente Division of Research, 2000 Broadway, Oakland CA 94612

Keywords

machine learning; risk prediction; critical care outcomes

Conventional wisdom suggests that we are climbing, or even cresting, the ‘peak of inflated expectations’ when it comes to Big Data and machine learning.(1) From this elevated perspective, the view is filled with the promise of peerless data and computation: opportunities to revolutionize the delivery of critical care. Unfortunately, common wisdom also suggests that what follows the mountaintop high is the ‘trough of disillusionment’, a period marked by the failure to deliver on over-hyped promises.(2) Despite the potential for changing sentiments about machine learning over the next few years, key contributions today will allow us to traverse and ultimately reach the ‘plateau of productivity’ in critical care.

In this issue of *Critical Care Medicine*, Weissman et al make several important contributions that help enhance that value for machine learning in critical care (3). Drawing from over 25,000 ICU episodes in the Medical Information Mart for Intensive Care (MIMIC) III data resource, they sought to develop a prediction model that could identify adult patients, starting early in the course of critical care, who would ultimately experience an adverse hospital outcome. Because their motivation was to develop tools that could be used to improve forecasting and prompt goals of care discussions, they aimed to predict the composite outcome of in-hospital death or an ICU length of stay of a week or longer.

Risk adjustment models that assess ICU patients’ likelihood of death or even length of stay are already highly developed in critical care.(4-6) Thus, the innovation in this study was two-fold. First, the authors extracted wholesale data from all clinical notes within the first 24 to 48 hours of admission and used a series of approaches to distill out the most valuable information contained within them. This allowed them to quantify the incremental contribution of key unstructured data (i.e., free text documentation) in prediction model performance. Second, they used a set of algorithms, with and without the benefit of the clinical documentation, to identify the machine learning method that yielded the best predictive performance.

Their key findings serve as important confirmation that advanced machine learning algorithms (i.e., gradient boosted trees) leveraging an expanded universe of data (i.e., clinical documentation) offer the highest predictive performance. For example, while a simple logistic regression model using only 18 standard variables like vitals and laboratory data exhibited very good discrimination (c-statistic: 0.79), a gradient boosted tree adding the 500 most essential terms from clinical documentation significantly enhanced performance (0.89). Their prior work suggests that other advanced machine learning algorithms, including neural networks similar to those recently used for mortality prediction models by Google(7), did not reliably improve model performance for this application.(8)

Perhaps an even more important contribution of this study is that the authors describe their approach in painstaking detail, both through extensive supplementary materials and an available code repository, ensuring that others interested in iterative improvements can start from a proven and high-performing baseline. This helps to avoid duplicative and proprietary efforts which limit the generalizability and utility of some tools. The open source approach is also highly aligned with widespread movements in machine learning and artificial intelligence that make cutting edge software and tools readily available to all users, rather than keeping them cloistered and inaccessible.

This study also highlights several challenges that will impact the value of machine learning tools in critical care. First, while the discrimination of the models improved with the use of advanced machine learning techniques, the incremental gains compared to a simple logistic regression were relatively modest. Using a workup-to-detection ratio framework (i.e., the number of patients reaching alert threshold who need to be evaluated to detect one case)(9), which has implications for the clinical implementation of a prediction model, the estimated differential between the simplest model (2.6 to 1) and the most complex model (1.7 to 1) may not have a major impact on reducing clinical burden. Second, of the 12 most important terms extracted from clinical documentation in a parsimonious model including only 25 variables, half were related to mechanical ventilation. Thus, the unstructured data appear to be capitalizing on clinical factors which are readily apparent to clinicians experienced with identifying patients at high risk for adverse outcomes. Interestingly, these high value terms differed from the authors' a priori terms of interest which largely revolved around prognosis. Third, the parsimonious model demonstrated very similar performance to more convoluted models, suggesting that simpler may be nearly as good as more complex. Finally, the discrimination exhibited by the advanced machine learning models, was similar to that described in contemporary iterations of standard ICU scoring systems for assessing hospital mortality.(10)

Despite the coming peaks and valleys of hype that will accompany the use of machine learning in critical care, rigorous, transparent, and important studies like this one will be essential to shorten the time it takes for us to reach the 'plateau of productivity'.

Acknowledgments

Funding: VXL is supported by NIH K23GM112018 and The Permanente Medical Group.

Copyright form disclosure: Dr. Liu's institution received funding from the National Institutes of Health (NIH)/ National Institute of General Medical Sciences; he received funding from the Permanente Medical Group; and he received support for article research from the NIH.

References

1. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med. 2017; 376(26):2507–2509. [PubMed: 28657867]
2. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA. 2017; 318(6):517–518. [PubMed: 28727867]
3. Weissman GE, Hubbard RA, Ungar LH, et al. Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay. Crit Care Med. 2018 in press.
4. Liu V. Keeping Score of Severity Scores: Taking the Next Step. Crit Care Med. 2016; 44(3):639–640. [PubMed: 26901551]
5. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care. 2010; 14(2):207. [PubMed: 20392287]
6. Verburg IW, Atashi A, Eslami S, et al. Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review. Crit Care Med. 2017; 45(2):e222–e231. [PubMed: 27768612]
7. Rajkomar, A., Oren, E., Chen, K., et al. Scalable and accurate deep learning for electronic health records. 2018. https://arxivorg/pdf/180107860pdf
8. Weissman GE, Hubbard RA, Ungar LH, et al. Inclusion Of Unstructured Text Data From Clinical Notes Improves Early Prediction Of Death Or Prolonged Icu Stay Among Hospitalized Patients. Am J Respir Crit Care Med. 2017; 195:A1084.
9. Romero-Brufau S, Huddleston JM, Escobar GJ, et al. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care. 2015; 19:285. [PubMed: 26268570]
10. Zimmerman JE, Kramer AA, McNair DS, et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today’s critically ill patients. Crit Care Med. 2006; 34(5):1297–1310. [PubMed: 16540951]