I Don’t Want My Algorithm to Die in a Paper
Detecting Deteriorating Patients Early

In this issue of the Journal, Pimentel and colleagues (pp. 44–52) report a retrospective evaluation of a new model (Hospital-wide Alerting via Electronic Noticeboard [HAVEN]) for predicting deteriorating ward patients using vital signs, laboratory measurements, demographics, and historical diagnostic coding (1). The standard metrics of accuracy are impressive (e.g., a c-statistic of 0.901). Accepting nine false alarms for every one true positive, HAVEN will identify more than 40% of cardiac arrests or unplanned ICU admissions within the preceding 48 hours and provide as much as 12-hour notice for more than 25%. This is twice the rate of the best of the alphabet soup of competitors (NEWS, LAPS-2, eCART, and several friends) (2–4).

But predictive scores with nice acronyms are two-a-penny. So why should we care? Because, reading the report carefully, this is a score that

References

1. Lim ZJ, Subramaniam A, Ponn_found R, Scott D, Jakes M, et al. Case fatality rates for patients with COVID-19 requiring invasive mechanical ventilation: a meta-analysis. Am J Respir Crit Care Med 2021;203:54–66.
2. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al.; Groupe de Recherche Clinique en REanimation et Soins intensifs du Patient en Insuffisance Respiratoire aiguE (GRC-RESPIRE) Sorbonne Université–Paris-Sorbonne ECMO-COVID investigators. Extracorporeal membrane oxygenation in severe acute respiratory distress syndrome associated with COVID-19: a retrospective cohort study. Lancet Respir Med 2020;8:1121–1131.
3. Shaeri S, Brenner SK, Gupta S, O’Garra BP, Krajewski ML, Charytan DM, et al.; STOP-COVID Investigators. Extracorporeal membrane oxygenation in patients with severe respiratory failure from COVID-19. Intensive Care Med 2021;47:206–221.
4. Diaz RA, Graf J, Zambrano JM, Ruiz C, Espinoza JA, Bravo SI, et al., National Advisory Commission for Adult ECMO. Extracorporeal membrane oxygenation for COVID-19–associated severe acute respiratory distress syndrome in Chile: a nationwide incidence and cohort study. Am J Respir Crit Care Med 2021;204:34–43.
5. Torres JP, O’Ryane L, Herve B, Espinoza R, Acuña G, Manalich J, et al. Impact of the novel influenza A (H1N1) during the 2009 autumn-winter season in a large hospital setting in Santiago, Chile. Clin Infect Dis 2010;50:860–868.
6. Schmidt M, Hajage D, Lepreton G, Monsel A, Voiriot G, Levy D, et al.; Groupe de Recherche Clinique en REanimation et Soins intensifs du Patient en Insuffisance Respiratoire aiguE (GRC-RESPIRE) Sorbonne Université–Paris-Sorbonne ECMO-COVID investigators. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome associated with COVID-19. A randomized controlled trial. JAMA 2020;324:1627–1639.
7. Schmidt M, Hajage D, Lepreton G, Monsel A, Voiriot G, Levy D, et al.; Groupe de Recherche Clinique en REanimation et Soins intensifs du Patient en Insuffisance Respiratoire aiguE (GRC-RESPIRE) Sorbonne Université–Paris-Sorbonne ECMO-COVID investigators. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome associated with COVID-19. A randomized controlled trial. JAMA 2020;324:1627–1639.
is built with implementation in mind. The authors state that “HAVEN recalculate(s) a patient’s deterioration risk each time a new variable is recorded” (1).

This is a substantially different challenge to that faced by the poster children of machine learning and artificial intelligence. Successful implementation of machine learning algorithms have largely focused on diagnostic imaging or pathways that do not require updating or real-time feedback loops (5, 6). In these, a digital image is captured, processed, and reported. The timescale follows the rhythm of the clinic visit. The data source is fixed and consistent.

In contrast, acute medicine does not have that luxurious timescale or reliable data source. Ward vital signs change as often as hourly, laboratory values are inconsistently measured, and the technology must update hundreds of patients instantly and simultaneously rather than batch process mammograms overnight.

Building a risk score that supports bedside clinical care for hundreds of patients across an entire hospital is as much about engineering as about machine learning. This requires a different set of trade-offs, and it is, therefore, exciting to see such a score performing well. Three criteria must be met to implement machine learning in such a scenario. These can be neatly mapped to the rapid response system reflex arc (7).

The afferent limb requires that data are digital first. Notably, the authors here have selected those items that are often digital even in hospitals that are only partway along the digital maturity spectrum (8). This suggests a broader impact than a model that depends on natural language processing of notes, current diagnostic information, or drug administrations. Moreover, the afferent limb requires a pipeline that delivers the digital signal to a computational engine near instantaneously. This is not part of the infrastructure provided by most electronic health record systems in which data is typically made available through a reporting data warehouse the following calendar day.

The second stage inserts a machine learning “synapse” before the efferent “effector” limb of the rapid response system. This must rapidly update the risk prediction in a trustworthy and reliable manner. Data entry errors or incomplete results must be managed (indeed the pattern of “missingness” might itself be informative). And “better” models that are computationally expensive must be eschewed for simple, “last one carried forward” median or mode approaches.

The third stage, the efferent limb, then returns this prediction to the bedside clinician in a manner that meets modern medical device engineering quality management standards (9).

Although the paper describes the development of the risk score, it has been done with these stages in mind. And despite the constraints this imposes, the score still performs. The closest similar work is the recent validation of a novel scoring system. Am J Respir Crit Care Med 2021; 204:44–52.

The prediction will then not serve these cohorts well. Such risks are mitigated by interpretable machine learning and by excellent and transparent reporting. Both are present here.

I now cannot wait to see the third stage being implemented and these tools returning predictions to the bedside. Work such as this is a stepping stone to true translational data science. We are not short of algorithms, but the majority of these end their lives in an academic report. For machine learning to have its promised impact on health care, we need engineering pipelines, and we need pragmatic algorithms. This paper hints at the former and strongly delivers the latter.

Author disclosures are available with the text of this article at www.atsjournals.org.

Steve Harris, Ph.D.
Critical Care Department
University College Hospital London,
London United Kingdom
and
Institute of Health Informatics
University College London
London, United Kingdom

ORCID ID: 0000-0002-4982-1374 (S.H.).

References

1. Pimentel MA, Redfern OC, Malycha J, Meredith P, Prytherch DR, Briggs J, et al. Detecting deteriorating patients in the hospital: development and validation of a novel scoring system. Am J Respir Crit Care Med 2021; 204:44–52.
2. Charpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards”. Crit Care Med 2014;42:841–848.
3. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013;84:465–470.
4. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. Med Care 2013;51:446–453.
5. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89–94.
6. Kim H-E, Kim HH, Han B-K, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020; 2:e138–e148.
7. Jones DA, DeVita MA, Bellomo R. Rapid-response teams. N Engl J Med 2011;365:139–146.
8. Cresswell K, Sheikh A, Krasuska M, Heeney C, Franklin BD, Lane W, et al. Reconceptualising the digital maturity of health systems. Lancet Digit Health 2019;1:e200–e201.
9. Guidance: medical device stand-alone software including apps (including IVDMDS). 2018 [accessed 20 May 2020]. Available from: https://www.gov.uk/government/publications/medical-devices-software-applications-apps.
10. Tomasiev N, Glorot X, Rae JW, Zielinski M, Ashkam H, Saraiya A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116–119.
11. Garroutte-Orgeas M, Boumendil A, Pateron D, Aegerter P, Somme D, Simon T, et al.; ICE-CUB Group. Selection of intensive care unit admission criteria for patients aged 80 years and over and compliance of emergency and intensive care unit physicians with the selected criteria: an observational, multicenter, prospective study. Crit Care Med 2009;37:2919–2928.