It doesn't matter what they say in the papers… It's still ROC and roll to me

William E. King[1]^, Brynne A. Sullivan[2]^, Zachary A. Vesoulis[3]^

[1]Medical Predictive Science Corporation, Charlottesville, VA, USA; [2]Department of Pediatrics, University of Virginia School of Medicine, Charlottesville, VA, USA; [3]Department of Pediatrics, Washington University in St. Louis School of Medicine, St. Louis, MO, USA

Correspondence to: William E. King. Medical Predictive Science Corporation, 1233 Cedars Court, Suite 201, Charlottesville, VA 22903, USA. Email: wking@heroscore.com.

Comment on: Chong SL, Niu C, Piragasam R, *et al.* Adding heart rate n-variability (HRnV) to clinical assessment potentially improves prediction of serious bacterial infections in young febrile infants at the emergency department: a prospective observational study. Ann Transl Med 2023;11:6.

Keywords: Heart rate variability (HRV); predictive analytics; early warning system (EWS)

Submitted Jan 16, 2023. Accepted for publication Feb 01, 2023. Published online Feb 09, 2023.
doi: 10.21037/atm-23-289
View this article at: https://dx.doi.org/10.21037/atm-23-289

Heart rate variability (HRV) is a quantitative method for calculating the variance in the length of time between successive heartbeats. Although healthy humans at rest are generally considered to have steady heart rates, there are minute fluctuations in the length of time between successive R waves, on the order of milliseconds. For more than half a century, a loss of HRV has been associated with autonomic dysfunction in the setting of illness. Perhaps the earliest description of HRV was by Hon and Lee in 1963 (1) with the observation that loss of HRV was associated with fetal distress and death, which formed the foundation for modern interpretative standards of cardiotocography (CTG) during labor.

In the decades that followed this observation, a wide range of HRV measures have been developed, evolving with advancements in computer technology to allow faster, more sophisticated calculations on increasingly large datasets. Alterations in HRV have been associated with increased risk of death after myocardial infarction (2), diabetes (3), depression (4), sepsis in adults (5) and neonates (6), and outcomes in neonatal hypoxic-ischemic encephalopathy (7). While many research tools for calculating HRV measures exist and the association of depressed HRV with illness is unequivocal, HRV monitoring is not a standard of care in the intensive care unit (ICU) or non-ICU setting.

Early warning of impending patient deterioration is an important and universal goal in health care. In contrast to doctors of yesteryear who had only a keen eye and a "sixth sense" at their disposal, the modern healthcare provider has vast sums of organized, digitized data at their fingertips. Big data predictive analytics hold great potential to improve patient health (8). However, this bounty has created a new problem-identifying salient and robustly predictive factors across heterogeneous and nuanced patient populations. A number of different systems have been developed to assess illness severity and predict morbidity or mortality. However, despite the initial promise of many of these systems, significant gaps have led to uneven adoption and failure to improve outcomes.

The Acute Physiology and Chronic Health Evaluation (APACHE) score was first proposed in 1985 by Knaus *et al.* (9) and encompasses lab values, limited vital signs, and acute and chronic medical conditions to provide a quantitative measure of illness severity and a prediction of mortality in the first 24 hours after admission to the ICU. Although currently in the fourth version (APACHE-IV), the

^ ORCID: William E. King, 0000-0002-4433-8797; Zachary A. Vesoulis, 0000-0001-8290-0069; Brynne A. Sullivan, 0000-0001-9580-4121.

APACHE-II score remains widely used due to ease of use and availability of online calculators. While performance is high for the first 24 hours, the APACHE score has been criticized for its static nature, failure to account for subsequent treatment or the hospital course after the first 24 hours, and interrater reliability concerns about certain variables (10,11).

The Sequential Organ Failure Assessment (SOFA) is a similar illness severity score designed with the explicit intent of quantifying the degree of organ failure in critical illness (such as sepsis) across cardiovascular, respiratory, neurologic, renal, hematologic, and hepatic domains (12). Like APACHE-II, SOFA is widely utilized in clinical practice. Pediatric (pSOFA) and neonatal (nSOFA) variants were developed more recently (13,14). Unlike APACHE-II, SOFA can be recalculated at regular intervals, accounting for longitudinal changes in the patient's course and treatment. The latest revision of consensus definitions of sepsis in adult patients uses SOFA as key criteria (15), yet significant concerns about SOFA's reliability and predictive performance have been raised (16,17).

Despite their utility for benchmarking ICU performance, research, or characterizing a septic patient, many argue that severity of illness scores like APACHE and SOFA should not be used for individual patient decision-making (18,19). To date, the literature lacks evidence demonstrating that clinical usage of SOFA leads to any change in patient outcome, despite widespread adoption.

Complimentary to early warning scores, has been the development and deployment of so-called "Rapid Response" teams (RRT), predesignated medical teams which are available to respond at short notice to medical wards when a decompensating patient is recognized. The RRT concept is based on several observations: (I) most in-hospital cardiac arrest occurs outside the ICU, (II) signs of decompensation are present 6–8 hours before arrest, and (III) similar dedicated teams responding to traumas led to improved morbidity and mortality outcomes. Although the implementation of RRT has been widespread, owing to recommendations from accreditation agencies, the quality of evidence supporting this practice remains low and inconsistent, as shown in more than one meta-analysis covering hundreds of studies (20,21).

While the application of Big Data analysis techniques may be technically possible and yield statistically impressive results, if it is not done with an understanding of the framework in which medicine is practiced, it is likely to struggle with real-world deployment, face significant adoption hurdles, and ultimately fail to positively alter patient outcomes.

Quantitative early warning algorithms should seek to make good on the promises of outcome calculators and rapid response teams by providing accurate, timely warning of an impending event where mitigation is likely to prevent damage or adverse outcomes. The challenge lies in integrating predictive analytics with clinician decisions (22). In doing so, such technology should reduce mortality and morbidity, but only if it performs as intended in the population to which it is applied. Failure to critically evaluate a predictive algorithm has resulted in many reportedly accurate algorithms falling short in the clinical setting. The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) guidelines were developed as a standardized reporting format to facilitate interpretation of results across studies (23). Here, we review six key principles to consider when evaluating and reporting the results of predictive analytics.

(I) Multiple metrics of accuracy should be evaluated and reported (24); for models developed using existing datasets where the outcome is known, the Area Under the Curve of the Receiver Operating Characteristic (Abbreviated AUC, AUROC, or ROC) is the most commonly used and an easily compared metric for performance evaluation. But AUC has drawbacks. First, it is calculated against every possible threshold of model output, and the tradeoff between sensitivity and specificity is calculated at each of these thresholds and summed. Yet, in clinical practice, the output of the model will result in distinct action thresholds (often only one action threshold, in fact). So, while AUC gives the modeler a wide picture of overall model performance, it does not measure the performance at the action threshold(s). Second, because AUC is derived from sensitivity and specificity, it is a backward-looking metric. Sensitivity and specificity start with a patient's known diagnosis and then look back to determine whether the test was positive for patients with the diagnosis (sensitivity) or negative for patients without the diagnosis (specificity). However, in clinical practice, the outcome is not yet known; thus, forward-looking metrics such as predicted risk, relative risk, odds ratio, positive predictive value, and negative predictive value better match the clinical picture confronting the

provider. Each of these metrics is calculated at a single threshold, which holds more clinically relevant information, but limits the range of predictive information conveyed. Calibration plots or measures provide information on how the accuracy of the predicted risk compared to the observed risk of an event or outcome. Importantly, a model with excellent discrimination (AUC >0.9) is rarely well-calibrated in the clinical setting, especially for rare outcomes.

(II) A second principle for consideration is the method and timing of action. Algorithms and early warning systems (EWSs) do not impact patients directly—they impact patients by prompting a change in clinical actions or decisions. The evaluator must ask two questions: (i) what will the predicted risk generated by the algorithm prompt the clinician to do that will intuitively affect patient outcome positively? And (ii) what is the desired timing relative to the event that an action (i.e., intervention) needs to occur in order to achieve the improved outcome? As a simple example that ties these concepts together, if we know that each hour of delay in starting antibiotics to treat sepsis is measurable in terms of increased mortality, then if we can improve the initiation of antibiotics by 3 hours with an EWS, we should be able to decrease mortality. Here we have defined an action (start antibiotics) that addresses the clinical event, and a defined time (3 hours) prior to the event at which to evaluate model performance.

(III) As alluded to above, a critical element of designing any EWS is implementation. A busy ICU is full of constantly alarming monitors, the vast majority of which are false positives and contribute to alarm fatigue. A new EWS should not further exacerbate this problem. Display of predictive analytics without alerts avoids this issue and allows users to interpret trends and risk in the context of the full clinical picture, but requires the clinician to seek out the information rather than having it automatically presented as an alert. Selecting an alert threshold that minimizes false positives will inherently decrease true positives, and may also alter the timing of the "early" warning to such a degree that there is no longer any action that can positively alter the outcome. Hence, the metrics of model performance must be calculated at the defined point in time relative to the clinical event (as outlined in the second principle). Continuing the sepsis prediction example, if we believe that antibiotics are over-utilized at a ratio of 10:1 in our patient population (which implies that standard care has a 9% positive predictive value), then we might choose a threshold that results in a 15% or 20% positive predictive value for early warning of sepsis, which would represent a dramatic improvement relative to the current standard of care.

(IV) Again, our previous principle has become a segue into the next: for any new predictive paradigm, quantitative comparison to the existing paradigm is necessary for determining if the newer, likely more complex, model has sufficiently greater performance to justify the increased computational and/or data input demands. Metrics of model performance reported for a new algorithm are contextual—an AUC of 0.70 or a PPV of 15% can benefit the patient if they represent improvements to the existing paradigm. Statistical tools such as the net reclassification index (25) quantify the degree to which performance shifts.

(V) Orthogonality provides a more complete picture of the patient. For example, in preterm infants, birth weight and gestational age provide nearly identical information, as opposed to gestational age and sex, which capture distinctly different elements of risk. Additionally, physiologic monitoring data stand to tell clinicians more about what they don't already know than other clinical risk factors. Many predictive algorithms use information that the provider is already aware of, in order to quantitate a risk prediction. This approach can only benefit the patient to the extent that the clinician is unaware or unable to process these data on their own. In contrast, algorithms that recognize subtle patterns in vital signs before the change is recognized at the bedside can alter the timing of interventions and change the course of illness. In practice, clinicians are remarkably adept at assimilating relevant clinical information, and through years of training and practice, intuiting the status of the patient. An algorithm that regurgitates back to the clinician something that they already know might provide decision support once the condition is recognized, but will not be useful for early warning. The patient will only benefit to the extent the model

is providing information that is either new (i.e., orthogonal) or beyond the capacity of the clinician to process in absence of the model.

(VI) Costs must be quantified. The deployment of any clinical decision support system involves costs, which are often far greater than the researcher initially conceives. Data acquisition interfaces, user interface hardware and software, regulatory and quality system burdens, installation and training costs, 24/7 technical support, and post-market surveillance are some of the many costs that must be borne. Cost-effectiveness for improving outcomes (e.g., reduction in length of stay, avoidance of complications or comorbidities) can be assessed against the capital or operating costs of the system or devices, where commercial options are available, or projected costs where such a system is envisioned. Costs should also be assessed along a different dimension—when predictions are inaccurate, what is the cost (financial, pain, increased length of stay) of additional testing and treatment for false positives? What is the cost (death, increased morbidities) for false negatives?

In a manuscript by Chong *et al.*, published of *Annals of Translational Medicine* (26), the authors describe a quantitative algorithm that uses conventional HRV, a novel second-order derivative of HRV (termed HRnV), demographics, vital signs, and laboratory results to identify which infants presenting to a hospital emergency ward were at risk for serious bacterial infection (SBI). In building this model, the authors utilized several of the previously outlined evaluation principles—the inclusion of orthogonal data elements and sequential comparison of more complex models using multiple accuracy metrics. The reported model shows significant promise for identifying infants at risk for SBI, permitting more targeted allocation of resources, and improving antibiotic stewardship, all of which are likely to increase the cost-effectiveness of care. We look forward to seeing future studies of this promising technology that evaluate thresholds for clinical action and external validation in a wider patient population. Given the financial and time costs of clinical trials, it is imperative that new models be rigorously evaluated and validated before deployment in a clinical setting or randomized clinical trials.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Annals of Translational Medicine*. The article did not undergo external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-23-289/coif). WEK is an employee, board member, and shareholder of Medical Predictive Science Corporation. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Hon EH, Lee ST. Electronic evaluation of the fetal heart rate. VIII. Patterns preceding fetal death, further observations. Am J Obstet Gynecol 1963;87:814-26.
2. Kleiger RE, Miller JP, Bigger JT Jr, et al. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. Am J Cardiol 1987;59:256-62.
3. Benichou T, Pereira B, Mermillod M, et al. Heart rate variability in type 2 diabetes mellitus: A systematic review and meta-analysis. PLoS One 2018;13:e0195166.
4. Carney RM, Blumenthal JA, Stein PK, et al. Depression, heart rate variability, and acute myocardial infarction. Circulation 2001;104:2024-8.
5. de Castilho FM, Ribeiro ALP, Nobre V, et al. Heart rate variability as predictor of mortality in sepsis: A systematic review. PLoS One 2018;13:e0203487.
6. Fairchild KD, O'Shea TM. Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis.

Clin Perinatol 2010;37:581-98.
7. Vergales BD, Zanelli SA, Matsumoto JA, et al. Depressed heart rate variability is associated with abnormal EEG, MRI, and death in neonates with hypoxic ischemic encephalopathy. Am J Perinatol 2014;31:855-62.
8. Vesoulis ZA, Husain AN, Cole FS. Improving child health through Big Data and data science. Pediatr Res 2022. [Epub ahead of print]. doi: 10.1038/s41390-022-02264-9.
9. Knaus WA, Draper EA, Wagner DP, et al. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818-29.
10. Polderman KH, Girbes AR, Thijs LG, et al. Accuracy and reliability of APACHE II scoring in two intensive care units Problems and pitfalls in the use of APACHE II and suggestions for improvement. Anaesthesia 2001;56:47-50.
11. Koperna T, Semmler D, Marian F. Risk stratification in emergency surgical patients: is the APACHE II score a reliable marker of physiological impairment? Arch Surg 2001;136:55-9.
12. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996;22:707-10.
13. Matics TJ, Sanchez-Pinto LN. Adaptation and Validation of a Pediatric Sequential Organ Failure Assessment Score and Evaluation of the Sepsis-3 Definitions in Critically Ill Children. JAMA Pediatr 2017;171:e172352.
14. Wynn JL, Polin RA. A neonatal sequential organ failure assessment score predicts mortality to late-onset sepsis in preterm very low birth weight infants. Pediatr Res 2020;88:85-90.
15. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315:801-10.
16. Kajdacsy-Balla Amaral AC, Andrade FM, Moreno R, et al. Use of the sequential organ failure assessment score as a severity score. Intensive Care Med 2005;31:243-9.
17. Zygun DA, Laupland KB, Fick GH, et al. Limited ability of SOFA and MOD scores to discriminate outcome: a prospective evaluation in 1,436 patients. Can J Anaesth 2005;52:302-8.
18. Soares M, Dongelmans DA. Why should we not use APACHE II for performance measurement and benchmarking? Rev Bras Ter Intensiva 2017;29:268-70.
19. Singer M, De Santis V, Vitale D, et al. Multiorgan failure is an adaptive, endocrine-mediated, metabolic response to overwhelming systemic inflammation. Lancet 2004;364:545-8.
20. Rocha HAL, Alcântara ACC, Rocha SGMO, et al. Effectiveness of rapid response teams in reducing intrahospital cardiac arrests and deaths: a systematic review and meta-analysis. Rev Bras Ter Intensiva 2018;30:366-75.
21. McGaughey J, Alderdice F, Fowler R, et al. Outreach and Early Warning Systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. Cochrane Database Syst Rev 2007;(3):CD005529.
22. Sullivan BA, Kausch SL, Fairchild KD. Artificial and human intelligence for early identification of neonatal sepsis. Pediatr Res 2023;93:350-6.
23. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63.
24. Harrell, FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer International Publishing; 2015.
25. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011;30:11-21.
26. Chong SL, Niu C, Piragasam R, et al. Adding heart rate n-variability (HRnV) to clinical assessment potentially improves prediction of serious bacterial infections in young febrile infants at the emergency department: a prospective observational study. Ann Transl Med 2023;11:6.

Cite this article as: King WE, Sullivan BA, Vesoulis ZA. It doesn't matter what they say in the papers… It's still ROC and roll to me. Ann Transl Med 2023;11(4):161. doi: 10.21037/atm-23-289

