American Journal of Biomedical Science & Research

www.biomedgrid.com

ISSN: 2642-1747

Mini Review

Lost in Translation? From Conventional Scoring Tools to Modern Data-Driven Risk Assessment in Critical Care Medicine

Martina A Maibach[1] and Jan Bartussek[1,2]*

[1]*Institute for Intensive Care Medicine, University and University Hospital Zurich, Switzerland*

[2]*Department of Quantitative Biomedicine, University and University Hospital Zurich, Switzerland*

***Corresponding author:** Jan Bartussek, Department of Quantitative Biomedicine & Institute for Intensive Care Medicine, University and University Hospital Zurich, Switzerland.

To Cite This Article: *Maibach A. M. and J. Bartussek. Lost in Translation? From Conventional Scoring Tools to Modern Data-Driven Risk Assessment in Critical Care Medicine. 2020 - 11(3). AJBSR.MS.ID.001626.* DOI: 10.34297/AJBSR.2020.11.001626.

Received: December 11, 2020; **Published:** December 16, 2020

Abstract

High-resolution, longitudinal health data became widely available in intensive care units in the past years. Patient risk assessment, however, is still primarily based on conventional scores that take into account only a few parameters taken at single time points, which frequently causes inaccurate predictions in the clinical practice. Likewise, the contribution of AI-approaches remains sparse, as current machine learning models are inherently difficult to deduce and even impressive results rarely contribute to disease understanding. This review focusses on the limitations of conventional risk scores, and on recent developments and challenges of novel, data-driven assessment tools.

Keywords: Critical Care Medicine, Risk Prediction, Decision Support, Machine learning, Artificial Intelligence

Introduction

Intensive care medicine is a fast-paced medical field where treatment decisions are often made ad hoc and without the full knowledge of the disease extend or patient background. In general, there is a slim margin for errors, and inaccurate or late treatment decisions have severe consequences and might even be fatal [1]. Therefore, a quick but precise assessment of the current patient status is crucial to reduce outcome morbidity and mortality [2]. In current medical practice, the evaluation of short-term disease progression is primarily based on clinical judgement of the treating physician [3,4]. A study by Fleig and colleagues demonstrated that decisions of physicians are highly subjective and often do not predict disease progression and outcome reliably [5]. Factors such as age, experience, attitude and religious as well as ethical beliefs of the physician influence their judgment [6-8]. Even under normal clinical conditions, health care professionals can cognitively assess only a limited amount of the continuous data stream offered by the monitors and machines of an intensive care set-up. Trends are therefore often recognized too late to allow targeted intervention. In the case of a crisis, the quality of human judgement might be further compromised by exhaustion, fear, the lack of current knowledge and scarceness of normally available resources [9]. Given the combination of high interpatient variability and the fast pace of critical care, standardized approaches are needed to classify patients into risk and progression subgroups that allow reliable assessments and prognosis.

Conventional Risk and Prognosis Scores Are Lacking

In order to objectify and assess highly complex and individual patient cases, a variety of risk assessment tools (scores) have been developed over the past 50 years [10]. Clinical scoring systems sum up points that are given if a physiological value or clinical parameter exceeds a defined threshold. For example, the APACHE II score integrates 12 physiological values, the age and chronic health problems to assess the mortality risk in intensive care patients [11]. Developed over 40 years ago, the APACHE II score is still widely used in clinical practice and research besides similar scores such as the SAPS II [12] or complementary scores such as the SOFA [13]. A common flaw of all scoring systems is that the achieved

score is compared to a historical patient group, often leading to an under- or overestimation of current mortality risks, especially when the current case is different from the case mixture of the original cohorts. For example, elderly patients (>80 years of age, high risk COVID-19 patients) are completely underrepresented in the cohorts of most established scores [14], severely compromising the accuracy of the prognosis in this patient group.

It is therefore not surprising that both the SAPS II and APACHE II score show a significant value for the Hosmer-learning-goodness-of-fit test ($p < 0.001$), indicating a poor mortality predicting performance, although both models showed moderately good AUC in the ROC analyses [14,15]. Another major limitation of existing scores is their lack of diagnostic value over time. Most intensive care risk assessment scores were developed as a one-time assessment within the first 24 hours of admission to the ICU and are not suitable for daily assessment. Nevertheless, these scores are used for repeated assessments due to a lack of alternatives. Moreover, scores do not take into account changes in medical conditions or medical resources and conventional prognostic scores can only predict the acute mortality risk at admission and do not take further endpoints, such as quality of live after ICU stay, into account. Furthermore, most general risk assessment scores do not differentiate between patient subpopulations that form due to the increasing modernization and specialization of medical fields. They perform poorly for specific diseases such as burn victims or cardio-chirurgic patients, but also for newly spreading diseases such as COVID-19 [16]. Especially in pandemic situations, the missing adaptability of scores is fatal due to lacking reliability in triage and resource allocation.

Artificial Intelligence for Medical Descision Support – Advances and Limitations

Continuous advances in medical technology and IT-infrastructure now provide access to high-resolution longitudinal patient data in unprecedented detail [17,18]. It is therefore not surprising that researcher have started to apply AI-approaches developed for big data analysis to create novel prognostic models. Especially the implementation of machine-learning techniques in critical care medicine has the potential to change the way medical progress is achieved significantly. A self-learning AI-system that supports treatment decisions through continuous analysis of all relevant patient data in real-time would be the "state-of-the-art" technology in tomorrow's critical care medicine. A prominent today's example is the early prediction of acute kidney injury, where an AI-system recognized a dangerous trend in the data 12-24h before the conventional laboratory test exceeded threshold [19]. Machine-learning has also been applied to the management of patients with viral respiratory infections [20] and the prediction of respiratory decompensation in the ICU [21]. However, utilizing artificial intelligence and especially machine-learning approaches on medical data has multiple drawbacks:

1. The prediction routines typically depend on the availability of a massive amount of training data (thousands of patients) [22]. In the case of novel or rare diseases, like COVID-19, such data is often not available. Moreover, the training and implementation of the underlying algorithms is computationally intensive and requires sophisticated IT-resources that are not commonly available.

2. Machine-learning is a stochastic, not a deterministic approach, meaning that it lacks common sense physical or physiological constraints [22,23]. Medical data collected during routine operation, however, is intrinsically flawed (i.e., missing, faulty or badly annotated data) introducing the possibility for logical fallacies in fully automated set-ups [22]. Furthermore, some multiple testing-based machine learning approaches utilizing massive amounts of data can run into the problem of P-hacking, i.e., finding random significant correlations due to the amount of tested correlations [24].

3. Machine-learning approaches generally have a poor transfer learning ability, meaning that each narrow machine-learning application needs to be specially trained. The results heavily depend on the selection and annotation of the training data that is often not representative for the general population. Therefore, many machine-learning approaches are too limited in scope and lack sufficient generalization to achieve a significant benefit for clinical routine [24,25]. Even if the training data is well selected, the same algorithm can arrive to different solutions performing well on the training set while acting very differently in real environments – a problem known as under specification [26]. Furthermore, ethical and moral constraints and concerns will make it difficult to translate any machine-learning-based solution that might affect human health and well-being [27,28].

4. The currently developed machine-learning algorithms might be able to predict a specific state reliably; however, a mechanistic understanding of the algorithm is often not possible, and no logical models can be derived from the final output [22,29]. Therefore, the final step in a machine-learning pipeline is often expert-based interpretation [30], which might be subjective and biased towards a certain hypothesis [31].

Conclusions and Future Perspective

The challenges encountered with machine-learning approaches are currently preventing a large-scale implementation of AI-systems in critical care medicine. Up to now, conventional risk and prediction scores are still the method of choice for decision support. A promising way forward would be to bring

back human knowledge into the training of machine-learning and other AI-models. The introduction of domain specific constraints, so called domain knowledge, may be the key factor missing [26]. This domain knowledge can be encoded in different ways: causal graphs [32], hybrid mechanistic-machine-learning models [33] and well-designed regularization schemes [34] just to mention some possibilities. This knowledge can act as a guide for the training of the algorithms, in order to choose the domain-relevant solutions from all the possible solutions it explores. Current research in these areas try to produce models that retain the predictive ability and performance that has fueled the current machine-learning boom, while achieving better generalization, robustness and some level of interpretability, all of these characteristics being essential in the medical domain.

Acknowledgements

This work has been financed by institutional fundings. We thank R. Schüpbach for his continued support.

Conflict of Interest

The Authors declare no conflict of interests.

References

1. Davis SS, WJ Babidge, GAJ Mc Culloch, GJ Maddern (2019) Fatal Flaws in Clinical Decision Making. ANZ J Surg 89(6): 764-768.
2. Gill TM (2012) The Central Role of Prognosis in Clinical Decision Making. JAMA 307(2): 199-200.
3. Westphal GA, AS Lino (2015) Systematic Screening Is Essential for Early Diagnosis of Severe Sepsis and Septic Shock. Rev Bras Ter Intensiva 27(2): 96-101.
4. Churpek MM, R Adhikari, DP Edelson (2016) The Value of Vital Sign Trends for Detecting Clinical Deterioration on the Wards. Resuscitation 102: 1-5.
5. Fleig V, F Brenck, M Wolff, M Weigand (2011) Scoring Systems in Intensive Care Medicine: Principles, Models, Application and Limits. Anaesthesist 60(10): 963-674.
6. Kong DF, KL Lee, FE Harrell, JM Boswick, DB Mark, et al. (1989) Clinical Experience and Predicting Survival in Coronary Disease. Arch Intern Med 149(5): 1177-1181.
7. Orav EJ, EA Wright, RH Palmer, JL Hargraves (1996) Issues of Variability and Bias Affecting Multisite Measurement of Quality of Care. Med Care 34(9 Suppl): SS87-SS101.
8. Kovacs G, P Croskerry (1999) Clinical Decision Making: An Emergency Medicine Perspective. Academic Emergency Medicine 6(9): 947-952.
9. Phillips-Wren G, M Adya (2020) Decision Making under Stress: The Role of Information Overload, Time Pressure, Complexity, and Uncertainty. Journal of Decision Systems 1-13.
10. Liao L, DMark (2003) Clinical Prediction Models. Journal of the American College of Cardiology 42: 851-853.
11. Knaus W A, E Draper, D Wagner, J Zimmermann (1985) Apache II: A Severity of Disease Classification System. Crit Care Med 13(10): 818-829.
12. Le Gall, J S Lemeshow, F Saulnier (1993) A New Simplified Acute Physiology Score (Saps II) Based on a European/North American Multicenter Study. Concepts in Emergency and Critical Care 270(24): 2957-3963
13. Vincent J L, R Moreno, J Takala, S Willatts, A DeMendona, et al. (1996) The Sofa (Sepsis. Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure. Intensive Care Med 22(7): 707-710.
14. Flaatten H, D de Lange, A Artigas, D Bin, R Moreno, et al. (2017) The Status of Intensive Care Medicine Research and a Future Agenda for Very Old Patients in the Icu. Intensive Care Med 43(9): 1319-1328.
15. Livingston B, F MacKirdy, J Hwie, J Ray, J Norrie, et al. (2000) Assessment of the Performance of Five Intensive Care Scoring Models within a Large Scottish Database. Crit Care Med 28(16): 1820-1827
16. Rapsang AG, DC Shyam (2014) Scoring Systems in the Intensive Care Unit: A Compendium. Indian J Crit Care Med 18(4): 220-228.
17. Torkamani A, KG Andersen, SR Steinhubl, EJ Topol (2017) High-Definition Medicine. Cell 170(5): 828-843.
18. Maibach M, A Allam, M Hilty, NA Perez Gonzalez, P Buehler, et al. (2020) The CoViD-19 ICU-Research Board Zurich, Timing Covid-19–Synchronization of Longitudinal Patient Data to the Underlying Disease Progression Using Crp as a Temporal Marker.
19. Tomasev N, X Glorot, J W Rae, M Zielinski, H Askham, et al. (2019) A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury. Nature 572: 116-119.
20. Mai M, M Krauthammer (2017) Controlling Testing Volume for Respiratory Viruses Using Machine Learning and Text Mining. AMIA Annu Symp Proc pp: 1910-1919.
21. Ren O, AEW Johnson, EP Lehman, M Komorowski, J Aboab, et al. (2018) Predicting and Understanding Unexpected Respiratory Decompensation in Critical Care Using Sparse and Heterogeneous Clinical Data. IEE Xplore pp: 144-51.
22. Ghassemi M, T Naumann, P Schulam, AL Beam, IY Chen, et al. (2019) A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc pp: 191-200.
23. Pearl J (2018) Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution. arXiv:1801.04016.
24. England JR, PM Cheng (2019) Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. AJR Am J Roentgenol 212(3): 513-519.
25. Dexter GP, SJ Grannis, BE Dixon, SN Kasthurirathne (2020) Generalization of Machine Learning Approaches to Identify Notifiable Conditions from a Statewide Health Information Exchange. AMIA Jt Summits Transl Sci Proc pp: 152-161.
26. D Amour A, K Heller, D Moldovan, B Adlam, B Alipanahi, et al. (2020) Underspecification Presents Challenges for Credibility in Modern Machine Learning arXiv:2011.03395.
27. Deo RC (2015) Machine Learning in Medicine. Circulation 132: 1920-1930.
28. Mc Cradden MD, S Joshi, M Mazwi, JA Anderson (2020) Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning. The Lancet Digital Health 2(5): E221-E223.
29. Knight SR, A Ho, Pius, I Buchan, G Carson, et al. (2020) Risk Stratification of Patients Admitted to Hospital with Covid-19 Using the Isaric Who Clinical Characterisation Protocol: Development and Validation of the 4c Mortality Score 370: m3339.

30. Rama K, H Canhao, AM Carvalho, S Vinga (2019) Aliclu - Temporal Sequence Alignment for Clustering Longitudinal Clinical Data. BMC Med Inform Decis Mak 19(1): 289.

31. Cabitza F, R Rasoini, GF Gensini (2017) Unintended Consequences of Machine Learning in Medicine. Jama 318(6): 517-518.

32. Stovitz SD, I Shrier (2019) Causal Inference for Clinicians. BMJ Evid Based Med 24(3): 109-112.

33. Suk HI, CY Wee, SW Lee, D Shen (2016) State-Space Model with Deep Learning for Functional Dynamics Estimation in Resting-State Fmri. Neuroimage 129: 292-307.

34. Webster KX, Wang, I Tenney, A Beutel, E Pitler, et al. (2020) Measuring and Reducing Gendered Correlations in Pre-Trained Models. arXiv:2010.06032.