Title: *Artificial intelligence in the clinical setting: Towards actual implementation of reliable outcome predictions*

Simon Tilma Vistisen[1,2]
Tom Joseph Pollard[3]
Steve Harris[4]
Simon Meyer Lauritsen[5,1]

[1] Institute of Clinical Medicine, Aarhus University, Denmark, vistisen@clin.au.dk, ORCID: 0000-0002-1297-1459
[2] Department of Anaesthesiology & Intensive Care, Aarhus University Hospital, Denmark
[3] Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA, tpollard@mit.edu, ORCID: 0000-0002-5676-7898
[4] Department of Critical Care, University College London Hospital and Institute of Health Informatics, University College London, UK, doc@steveharris.me, ORCID 0000-0002-4982-1374,
[5] Enversion A/S, Denmark sla@enversion.dk, ORCID: 0000-0001-8823-5047

Communicating author:

Simon Tilma Vistisen
Email: vistisen@clin.au.dk
Associate Professor, Aarhus University Hospitals

Palle Juul Jensens Boulevard 99, C319-128, 8200 Aarhus N, Denmark

+45 2067 6868 (Cell phone)

Word count, body: 1725
Number of tables: 1
Number of figures: 0

Key words: Artificial intelligence, machine learning, implementation, framing

Acknowledgements relating to this article

1. Assistance with the article: None
2. Financial support and sponsorship: None
3. Conflicts of interest: None.

Advanced statistical models for predicting adverse clinical events have become omnipresent in the literature and we often hear that concepts like *artificial intelligence* or *machine learning (ML)* are going to disrupt medicine. Given the amount of data generated during surgical procedures and intensive care admissions, these clinical areas are prototypical for the application of ML. Yet, in the face of massive attention and enormous research output, there are so far few clinically validated and implemented algorithms.[1] Within the anaesthesia and intensive care area, we are familiar with few compelling sepsis prediction studies, yet they are either small[2] or not designed as a randomised controlled trial.[3] In this editorial, we broadly discuss some of the reasons why ML struggles with real-world implementation. Some of these reasons relate to *methodology*, others to *clinical context*.

Framing the question

Few machine learning researchers are intimately familiar with the clinical environment, and so it should be no surprise that many machine learning studies are not carried out in a way that allows for easy translation to the bedside. Framing a machine learning study appropriately – that is, properly defining the clinical event and the prediction task – requires interdisciplinary knowledge and detailed discussion of methodology. For a prediction task, for example, framing would include identifying the clinical outcome, specifying when exactly the prediction is made, selecting the observation window, and so on. These details are sometimes poorly considered, sometimes poorly described. Framing forms the very backbone of the machine learning model being developed, and evaluation takes place within the context of the framing.[4] Consequently, without clear and clinically relevant framing a seemingly high-performing model may still not be clinically usable.[5] Many machine learning studies seek to address clinically relevant problems, but oversimplify the problem to the point where clinical relevance is eventually lost. The ubiquitous case-control framing/design in machine learning studies is a good example of designs where researchers seek to solve a clinically relevant problem, which is not aligned with clinical reality. The evidence level of a classic case-control study is weak and the caveats of this design, such as selection bias, does not disappear just because a study applies machine learning techniques. In relation to creating models that can make predictions and update them over time, applying the case-control design in a “validation study” is often creating a *temporal* bias that should be avoided.[6] When releasing a black box prediction algorithm that is developed this way, the result is often that the positive predictive value declines dramatically[6] and that it is impossible for users to know which event alarms to trust.

The nature of observational data

Many studies are based on analyses of large retrospectively collected datasets, where *missing data* is a frequent and natural phenomenon. The treatment of missing data is often a major issue, given that data is rarely missing at random. One could think of the simple *physiological* example of SpO_2 becoming unmeasurable in shock/hypotension. A *clinical* example is the difference between the patient who had an arterial blood gas taken in the emergency department (ED) versus the patient who did not. A *clinician* decided to obtain that blood gas. This presence or missingness of an observation tells us something important. Taking this a step further: Where and when was the blood gas taken? If taken in the first postoperative hours in the cardiac surgery recovery unit, that lab test result could well be obtained to inform FiO_2 adjustment, indicating a different “lab presence risk” than in the ED patient. A large retrospective study found that the mere “presence of a laboratory test order, regardless of any other information about the test result, has a significant association with

the odds of survival in 233 of 272 (86%) tests. Data about the timing of when laboratory tests were ordered were more accurate than the test results in predicting survival in 118 of 174 tests (68%)."[7] Observational studies, whether retrospective or prospective, are in general prone to this missingness bias. While imputation techniques and use of auxiliary variables may help to mitigate these issues, there should be no expectations of generalisability given that missingness patterns likely reflect a specific ward's clinical culture.[8,9]

Algorithm performance

In machine learning research, often the goal is to outperform previous literature on benchmark tasks, rather than to truly consider how models might perform in practice. Many studies that use large datasets are focused on the applied methods and algorithms,[10] as well as fine-tuning of performance metrics such as area under the receiver operating characteristics curve (AUROC), sensitivity and specificity. This focus on algorithms and classification performance often comes at the expense of basic epidemiologic principles and clinical interpretations[11] and it is characterised by vague basic descriptions of the data and its origin, often indicating a limited domain knowledge among authors.

Oversimplified Methods sections and lack of code sharing often result in the inability of the community to even reproduce a study cohort, let alone the study outcome.[12] The unfortunate consequence is that algorithms become useless to the research community, contributing to research waste.[10]

Fortunately, the unfavourable events that are typically the goal of prediction tasks occur rarely. Metrics such as AUROC, sensitivity and specificity are important features of an algorithm, but they do not address *clinical* usefulness and may be less informative in cases of rare events. For "imbalanced" datasets, a metric such as the area under the precision-recall (PR) curve - plotting the positive predictive value (precision) against sensitivity (recall) - is often more informative, particularly in order to quantify the presence of false alarms. For the clinician at the bedside, a model with a high false alarm rate is unlikely to be a useful model. In addition, if a false positive decision causes greater harm than a false negative decision, a model with high specificity may be preferable to a model with high sensitivity and lower specificity, although the latter model might have, say, a higher AUROC. In general terms, a model is clinically useful if the use of its decisions for patients leads to a better ratio between benefits and harms than not using the model.[11,13]

The decision for converting a predicted probability into a binary label (positive or negative) is governed by a decision threshold in the range between 0 or 1. For example, with a decision threshold of 0.5, probabilities less than 0.5 are assigned to class 0 and values greater than or equal to 0.5 are assigned to class 1. ROC and PR curves are all diagnostic plots that evaluates a set of probability predictions at varying decision threshold. In the case of a ROC curve, a set of different thresholds are used to interpret the true positive rate and the false positive rate of the predictions on the positive. In this sense, the ROC is a useful tool to understand the trade-off in the true-positive rate and false-positive rate for different thresholds.

Similarly, decision curve analysis (DCA)[14] assesses the clinical usefulness of a prediction model by evaluating the so-called net benefit at varying decision thresholds for the model. In practice, this means that the decision threshold is used to control the exchange ratio between the number of false positives that is acceptable in exchange for one true positive. This interpretation is important, because it is informative of how the clinician weights the harm of a false decision over the benefit of a true decision. The harm/benefit exchange ratio is subjective and will vary across clinicians. A decision curve in DCA illustrates the

consequence of an arbitrary choice by evaluating the net benefit for the binary decision of opting into the intervention or not across a range of different decision thresholds – or equivalently, for a range of different harm-benefit exchange ratios.[15]
Another key aspect of model performance that is often overlooked is algorithmic bias. Does the model exhibit behaviour that might reinforce inequalities? Strong overall performance of a model can be misleading, concealing poor performance in patient subgroups.

Validation and trust

If a new model is being proposed then there is almost no reason not to provide one or more reference models for comparison. These might include a classic regression model and clinical scores of disease severity (e.g. APACHE, SOFA, etc.). Yet, reference models are often missing, which makes it impossible to determine if a new and less transparent model is adding any predictive value (at the expense of direct interpretability).[16] In low-risk-of-bias studies, where an interpretable logistic regression model *is* reported, more advanced machine learning models rarely outperform logistic regression.[17] Methodological issues such as this may become less common once reporting guidelines are established for diagnostic and prognostic prediction model studies based on artificial intelligence,[10] and their subsequent uptake into the reviewing process of scientific journals. While developing approaches that enable the reasoning of complex machine learning models to be explained is an active research area, it is fair to say that this is still in its infancy. Interpretable models are likely to be preferred by clinical teams, even at the expense of performance, favouring traditional modelling approaches over the “black box” of state-of-the-art models such as neural networks.

Deployment at the bedside

There are also crucial *contextual* and *technical* reasons why so few artificial intelligence algorithms (even well-validated ones from a scientific point of view) have been deployed successfully.[18] Arguably, studies that explore translation of algorithms to the bedside are scarce, at least in part because the academic system provides greater reward to the lower-hanging fruits of fast, successive publications that are unencumbered by the realities of the clinical environment. Technically, there is a huge gap between the data infrastructure needed to train an algorithm on a retrospective dataset, extracted once from a setup optimised for collecting and storing data, and using the algorithm in a prospective, and maybe real-time, setup. There is also the issue of dataset shift. The clinical environment and its patient population is not static. A model that works now may catastrophically fail when a laboratory reagent is switched, a protocol is updated, or patient demographics change.

In conclusion, a number of challenges, summarised in Table 1, have inhibited widespread adoption of machine learning models at the bedside, but there has been progress nonetheless. This progress includes movement towards collaborative approaches for machine learning in health research; public datasets that are more representative of the clinical environment; more holistic metrics for assessing performance; and establishment of guidelines for reporting machine learning studies.

References

1. Ben-Israel D, Jacobs WB, Casha S, *et al.* The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine* 2020; **103**, **10178**.
2. Shimabukuro DW, Barton CW, Feldman MD, *et al.* Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. *BMJ Open Respiratory Research* 2017; **4**:e000234.
3. Escobar GJ, Liu VX, Schuler A, *et al.* Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *New England Journal of Medicine* 2020; **383**:1951–1960.
4. Kanter JM, Gillespie O, Veeramachaneni K. Label, Segment, Featurize: a cross domain framework for prediction engineering. *2016 IEEE International Conference on Data Science and Advanced Analytics* 2016:430–439.
5. Lauritsen SM, Thiesson B, Jørgensen MJ, *et al.* The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. *npj Digital Medicine* 2021; **4**:1–12.
6. Yuan W, Beaulieu-Jones BK, Yu KH, *et al.* Temporal bias in case-control design: preventing reliable predictions of the future. *Nature Communications* 2021; **12**:1–10.
7. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ (Online)* 2018; **361**:1479.
8. Li J, Yan XS, Chaudhary D, *et al.* Imputation of missing values for electronic health record laboratory data. *npj Digital Medicine* 2021; **4**:1–14.
9. Kyono T, Zhang Y, Bellot A, *et al.* MIRACLE: Causally-Aware Imputation via Learning Missing Data Mechanisms. 2021.
10. Collins GS, Dhiman P, Andaur Navarro CL, *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021; **11**:48008.
11. Shah NH, Milstein A, Bagley Steven C. P. Making Machine Learning Models Clinically Useful. *JAMA* 2019; **322**:1351–1352.
12. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. *Proceedings of the 2nd Machine Learning for Healthcare Conference* 2017:361–376.
13. Vasey B, Clifton DA, Collins GS, *et al.* DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine* 2021; **27**:186–187.
14. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research* 2019; **3**:18.
15. Zhang Z, Rousson V, Lee W-C, *et al.* Big-data Clinical Trial Column Decision curve analysis: a technical note. *Ann Transl Med* 2018; **6(15)**:1–11.
16. Vistisen ST, Alistair ·, Johnson EW, *et al.* Predicting vital sign deterioration with artificial intelligence or machine learning. *Journal of Clinical Monitoring and Computing* 2019; **33**:949–951.
17. Christodoulou E, Ma J, Collins GS, *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 2019; **110**:12–22.
18. Kelly CJ, Karthikesalingam A, Suleyman M, *et al.* Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* 2019; **17**:195.

Table 1: Issues and challenges often present in the exiting prediction model studies and possible ways to handle them

Issues to address or handle	**Possible ways to handle the issues in future studies**
Insufficient reporting of design and framing	Always align the framing and design with the clinical question/unmet need The classic case-control design is rarely aligned with a clinically relevant question. Report details about observation window, prediction window, lead time, window shift, preferably with a supporting figure
Missing values	Quantify the presence and its implications Possibly apply imputation if deemed meaningful Possibly model the missingness pattern to inform the prediction model
Insufficient reporting and clinical assessment of discrimination metrics	Thoroughly report discrimination metrics Discuss the presence and implications of a possibly unbalanced design/dataset Discuss perceived clinical benefits of the prediction model (compared with the reference model), e.g. using concepts of net benefit and decision curve analysis
Lack of a clinically meaningful reference model	Always report a reference model that could be considered current practice for predictions. Example of short-term outcomes: Predicting eminent hypotension or tachycardia: Blood pressure or heart rate itself, respectively, should always be (part of) a reference model. Preferably a transparent regression model if multivariate. Example of Longer-term outcomes: Predicting sepsis or mortality: Clinical scores of disease severity, such as EWS, SOFA or APACHE scores, can be relevant or a multivariate regression model based on the underlying variables used for such scores in order to calibrate better
Obstacles for actual implementation	Discuss why and where the prediction model could realistically be implemented