Machine learning in intensive care medicine: ready for take-off?

Lucas M. Fleuren, MD (1,2)
Patrick Thoral, MD (1)
Duncan Shillan, MD (3)
Ari Ercole, MD, PhD (4,5)
Paul W.G. Elbers, MD, PhD, EDIC (1,5)
On behalf of the Right Data Right Now collaborators

Right Data Right Now Collaborators
Mark Hoogendoorn, PhD (2)
Ben Gibbison, MD (3)
Thomas L.T. Klausch, PhD (6)
Tingjie Guo, MSc (1)
Luca F. Roggeveen, MD (1,2)
Eleonora L. Swart, PhD (7)
Armand R.J. Girbes, MD, PhD, EDIC (1)

(1) Department of Intensive Care Medicine, Research VUmc Intensive Care (REVIVE), Amsterdam Medical Data Science (AMDS), Amsterdam Cardiovascular Sciences (ACS), Amsterdam Infection and Immunity Institute (AI&II), Amsterdam UMC, location VUmc, VU Amsterdam, The Netherlands

(2) Computational Intelligence Group, Department of Computer Science, VU Amsterdam, The Netherlands

(3) NIHR Bristol Biomedical Research Centre, University of Bristol, Bristol, UK

(4) Division of Anaesthesia, University of Cambridge, Cambridge, United Kingdom

(5) Data Science Section, European Society of Intensive Care Medicine

(6) Department of Epidemiology and Biostatistics, Amsterdam UMC, location VUmc, VU Amsterdam, The Netherlands

(7) Department of Pharmacy, Amsterdam UMC, location VUmc, VU Amsterdam, The Netherlands

In 1986 the world was shaken by the Challenger space shuttle disaster. In the years that followed, the American National Aeronautics and Space Administration (NASA) called for a strategy change in space technology development [1]. Allowing technology to be developed without a specific space program in mind was central to the new strategy [2]. In order to evaluate resulting projects with no direct contribution to a space mission, NASA introduced the general concept of technology readiness levels (TRLs) [3]. These nine levels, adopted by many EU institutions, assess the maturity level of technology and estimate its readiness to fly.

As machine learning is taking flight in the medical domain, intensive care medicine is facing a similar evaluation problem. Despite a surge in innovative models trained on intensive care data[4], it remains unclear which projects could actually make it to the patient's bedside and improve care. We hypothesize that machine learning projects follow a trajectory to the patient's bedside analogous to the way aerospace technology ventures into outer space. Therefore, we set out to translate the NASA technology readiness levels into a clinical applicable scale. We consequently applied the scale to ICU machine learning literature.

A panel of three experienced intensivists in medical data science research (PT, AE, PE) and an associate professor in machine learning (MH) iterated translations of the NASA TRLs into a clinically applicable scale until all unanimously agreed (see Table 1). Three authors (AE, PE, LF) applied the scale to all critical care machine learning papers identified by Shillan et al. in their recent review[4], where each paper was reviewed by at least one intensivist. Articles published before 2008 (n=55), pediatric articles (n=27), and reviews (n=2) were excluded. After an initial random 20 papers were reviewed, all panel members agreed level 3 and 4 be merged into a single 'model development' level. Any discrepancies in the final scoring were adjudicated by two panel members (PE, LF).

The clinical readiness levels for machine learning is presented in Table 1. A total 172 articles were scored, of which 160 articles (93%) scored level 4 or below, 8 articles (5%) validated results on data other than the initial data-split, 2 articles (1%) integrated a model into the workflow without exposing clinicians to the results, and only 2 articles (1%) evaluated models against clinical relevant outcomes. Reports on model integration (level 9) were not found.

No single study design is suited to evaluate all departures of machine learning models into the clinical workflow, and some warrant more extensive testing than others. However, demonstration of model safety and efficacy is paramount for the transition from bench to bedside and to convince clinicians of potential benefits. Although a limitation of the study is that only models found in medical literature were considered, we would expect adequately designed and tested models to be published. The small minority of clinically implemented projects identified here arguably form a large gap to be bridged between bytes and bedside. With the current framework we hope to encourage critical appraisal of machine learning projects in order to estimate their readiness to fly.

Bibliography

1. Héder M (2017) From NASA to EU: The evolution of the TRL scale in Public Sector Innovation. Innov J 22:1–23

2. Sadin SR, Povinelli FP, Rosen R (1989) The NASA technology push towards future space mission systems. Acta Astronaut 20:73–77. https://doi.org/10.1016/0094-5765(89)90054-4

3. Mankins JC (1995) Technology readiness levels

4. Shillan D, Sterne JAC, Champneys A, Gibbison B (2019) Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. Crit Care 23:1–11. https://doi.org/10.1186/s13054-019-2564-9

TRL	NASA Definition	Clinical Definition	Explanation	Examples	Papers
1	Basic principles observed and reported	Clinical problem identification	Identification of potential medical machine learning solution to defined clinical problem. Appraisal of the literature is carried out. Medical research question is defined	Review of the literature giving clinical motivation for model Detailed clinical problem definition with identification of gap for novel machine learning solutions	0%
2	Technology concept and/or application formulated	Proposal of model/solution	Elaborate proposal of potential medical machine learning solution to clinical problem. Medical dataset identified and outcomes defined	Project proposal with Detailed description of proposed methods Proposal registration prior to development	0%
3	Analytical and experimental critical function and/or characteristic proof-of-concept	Model prototyping & Model development	Prototype model or model development. Demonstrating potential for prediction or decision support or optimizing model and validation on characterized data	Feasibility studies of novel machine learning techniques on medical data Demonstration that data contains predictive information	93%
4	Component and/or breadboard validation in laboratory environment			Study of machine learning model performance on prespecified clinical task with internal clinical data or online datasource	
5	Component and/or breadboard validation in a relevant environment	Model validation	Representative model is validated on realistic dataset other than original training and testing population (ideally data pre-processing and cleaning pipelines are specified for all data sources)	Model validation study on retrospective data from other centers Model validation study on prospectively collected data, not real-time Studies reporting development of model and external validation	5%
6	System/subsystem model or prototype demonstration in a relevant environment	Real-time model testing	Model performance is tested real-time and integrated into the EHR/hospital system in one or more clinical settings, but no implementation into clinical workflow (i.e. no clinical staff is exposed to model results)	Prospective observational studies E.g. comparing model performance to standard of care Technical pipelines for automated clinical data extraction demonstrated as appropriate	1%
7	System prototype demonstration in an operational environment	Workflow implementation	Model is implemented in clinical workflow (i.e. clinical staff is exposed to model results). Performance, safety, and effects on patient relevant outcomes are assessed. Workability by clinical staff is evaluated. 'Phase 2' Study	Prospective 'phase 2' safety / usability	0%
8	Actual system completed and qualified through test and demonstration	Clinical outcome evaluation	Implemented model evaluation against clinical outcomes. 'Phase 3' study.	Randomized clinical intervention trial implementing the model and comparing clinically relevant outcomes to a control group Pre-post implementation intervention studies comparing clinically relevant outcomes	1%
9	Actual system proven through successful mission operations	Model integration	Model has proven to work in a research setting in its final form and in expected conditions. Model is integrated in the clinical workflow and evaluated in different centers (if appropriate) whether it meets specifications	Long term model evaluation after integration in a clinical workflow Studies reporting changes in clinical care and adaptation to model by clinical personnel Guidelines recommending machine learning tools in routine clinical practice Post-implementation surveillance studies	0%

Table 1. Clinical machine learning readiness levels. TRL = technology readiness level, NASA = American National Aeronautics and Space Administration