The challenge of implementing AI models in the ICU

Life In View/Science Photo Library

Published Online November 8, 2018
http://dx.doi.org/10.1016/S2213-2600(18)30412-0

See Articles page 905

The rapid emergence of artificial intelligence (AI) deep learning models in health care has generated expectations that AI and its ability to deal with huge, complex, rapidly updating data arrays will deliver better patient care in data-rich intensive care units (ICUs), with such models holding the potential to inform decision making, alongside complementary methodological advances in modelling of ICU outcomes.[1] In *The Lancet Respiratory Medicine*, Alexander Meyer and colleagues report promising findings from deep learning recurrent neural network (RNN) models predicting severe complications in real time in patients recovering from cardiac surgery in intensive care.[2] The authors considered three outcomes—mortality, renal failure with a need for renal replacement therapy, and postoperative bleeding leading to operative revision—and compared the predictive quality of their models against established standard-of-care clinical reference tools (Bojar's algorithm for postoperative bleeding, the Simplified Acute Physiology Score II [SAPS II] for mortality, and the Kidney Disease: Improving Global Outcomes staging criteria for acute renal failure). The RNN models were trained and tested on 11 492 intensive care admissions, corresponding to 9269 patients, and the results validated externally with cases from a published dataset. With 52 patient features included in the models, 39 of which were dynamic and so could substantially change during hospitalisation (eg, blood pressure), the authors acknowledge that the volume and complexity of ICU data are already beyond human processing but are ideally suited for AI and its deep learning methods.

Given recent developments in these methods, it is not surprising that Meyer and colleagues' models perform well statistically. The authors assessed the performance of their RNN models with a series of statistical measures, including area under the curve (AUC), which is a measure of how well a model can distinguish between classes; in this case, the two classes were whether a patient was going to experience a severe complication or not. Mortality prediction improved the AUC by 0·24, from the 0·71 achieved by SAPS II to 0·95 achieved by the RNN model. Similarly, prediction of bleeding with the RNN model increased the AUC by 0·29 from the 0·58 achieved by Bojar's algorithm to 0·87. For individual prediction, the positive predictive value (PPV) of the RNN model for bleeding was 0·84, with a negative predictive value (NPV) of 0·77. Translating these numbers to the clinic, a PPV of 0·84 for bleeding indicates that 16 of 100 patients predicted to bleed are false positives, risking needless treatment; likewise, an NPV of 0·77 tells us that 23 of 100 patients who are predicted not to bleed are false negatives, thus risking missed interventions. Prediction of mortality with the RNN model saw a slightly better performance, with a PPV of 0·90 and an NPV of 0·86, meaning ten of 100 patients predicted to die would be false positives and 14 of 100 predicted to survive would be false negatives.

These predictive results are important. The authors rehearse the model's improvement over the clinical reference tools. However, by doing so, they miss the real target. A clinical algorithm for bleeding with an AUC of 0·58 is not performing much better than random guessing (for which the AUC would be 0·50), so few clinicians would use such an algorithm with any conviction, and thus it should not provide the benchmark for a model's success. The promise of high performance and robust signals of RNN models will generate great user expectations and confidence, in contrast with fallible simple clinical guides; this promise, however, comes with the caveat that the workings and predictions of these models cannot be easily checked or adjusted. Therefore, the validated statistical performance of these RNN models needs to be very high to justify the trust clinicians will have to put in them. PPVs and NPVs of 0·77–0·84 are perhaps not sufficiently convincing in this context.

Meyer and colleagues' models use raw, uncurated data without requiring time-consuming querying or correction, which would prohibit true real-time predictions. However, data quality is important because "Bad data can be amplified into worse models".[3] A specific concern about the data quality of electronic health records is misclassification and mismeasurement.[4] Advantages of RNN models[5,6] are that they continuously learn, and as data feeds get larger and richer (eg, by including genomics and imaging data), computing power increases, and the methodological engine improves, so the statistical performance of such models will improve further. Training RNNs on local data in both space and time allows the models to be personalised to diverse sets of patients in diverse health-care settings, and hence, in time, their application might have the power to decrease health inequalities globally. Meyer and colleagues' RNN models also hold up reasonably well in the validation cohort,[2] which is important for reliable implementation.

However, implementation is one of the most difficult challenges to realising the benefits of these decision-support tools. Meyer and colleagues acknowledge that models need testing prospectively,[2] recognising that where models predict outcomes, clinicians interpret signs, make decisions, and change behaviours to generate benefit for the patient.[7] The clinician's use of models for enhanced decision making needs rigorous assessment: do they understand outputs and interpret them correctly, and does model use lead to measurable changes in decisions that generate measurable patient benefits? The authors suggest that models could be directly integrated into existing electronic health record systems; however, this underestimates the substantial practical challenges to ensuring decision-support tools can process real-time data feeds and robustly deliver continuous guidance in a user-friendly format within the ICU. Industrial-academic collaborations (eg, IBM Watson and Google DeepMind) have encountered substantial problems in this regard.[8] In addition, issues arise around data ownership and access, given possible commercial exploitation of the required data feed inputs. [9]

The greatest challenge will be integrating these decision-support tools into multidisciplinary shared decision making, ensuring that all teams involved are comfortable with an increased reliance on decision aids;[10] a further challenge is making sure that patients and carers are properly informed about the methods being used and approve of this increased reliance on data-driven clinical decision making. This communication is difficult, not least because, unlike models such as the simple logistic prediction model that works with easily understood factors (eg, age, gender, or APACHE score), deep learning rests upon hidden factors that uncover complex, high-order data patterns; by definition, we do not know—and far less are able to describe or explain—what such a model is doing.

Any implementation of such models requires flawless operating performance. This real-world performance needs to be convincingly and continuously demonstrated. For example, how robust are predictions if crucial data feeds, alone or in combination, go offline? What happens if a crucial data feed goes out of calibration, degrading in quality until it potentially becomes just noise? Would the model be able to learn internally if or when its predictions were becoming unacceptably accurate and switch itself off?

Many believe widespread deployment of these decision-support tools is inevitable and will generate better effectiveness and increased safety. For example, better decisions around which drugs to prescribe at which time and dose should decrease prescribing errors. Additionally, Meyer and colleagues' model predicts future complications to initiate early interventions, which are assumed proven to be safe, effective, acceptable, and affordable. However, it could be that future models will address more vexed questions of withholding treatments (eg, on the grounds of futility or cost), generating complex clinical, ethical, and—inevitably—legal issues.

Meyer and colleagues' deserve congratulations on an elegant Article,[2] confirming that RNN models in demanding patient groups and settings are within our grasp. However, the effective, safe, and affordable implementation of these models present a vast number of challenges. The excellent statistical performance of such tools are likely to improve further—and quickly. The remaining challenges, such as the human factors of training the practitioners, redesigning service delivery to integrate these tools, and bringing the patients fully on board in this adventure, are more difficult to overcome and doing so will require diverse skills and disciplines. Ongoing rigorous evaluation of benefits, safety, acceptability, and costs[8] is necessary, perhaps using hybrid implementation-effectiveness designs—which allow the integrated evaluation of both the effectiveness of an intervention while addressing the facilitators and barriers to implementation.[11] These challenges must be faced and overcome if we are to realise the undoubted potential benefits of AI in health care.

John Norrie
Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh, Edinburgh, EH16 4UX
j.norrie@ed.ac.uk

I declare no competing interests.

1 Norrie J. Mortality prediction in ICU: a methodological advance. *Lancet Respir Med* 2015; **3:** 6–7.

2 Meyer A, Zverinski D, Prahringer B, et al. Machine learning for real-time prediction of complications in critical care: a rerospective study. *Lancet Respir Med* 2018; **6:** 905–14.

3 Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA* 2018; **319:** 19–20.

4 Mullainathan S, Obermeyer Z. Does machine learning automate moral hazard and error? *Am Econ Rev* 2017; **107:** 476–80.

5 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.

6 Hinton G. Deep learning – a technology with the potential to transform health care. *JAMA* 2018; **320:** 1101–102.

7 Obermeyer Z, Lee TH. Lost in thought—the limits of the human mind & the future of medicine. *N Engl J Med* 2017; **377:** 1209–11.

8 Lancet. Artificial intelligence in health care: within touching distance. *Lancet* 2017; **390:** 2739.

9 Naylor CD. On the prospects for a (deep) learning health care system. *JAMA* 2018; **320:** 1099–100.

10 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; **318:** 517–18.

11 Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care* 2012; **50:** 217–26.

A new low represents a new high in surgical safety

CC Studio/Science Photo Library

Published **Online**
November 12, 2018
http://dx.doi.org/10.1016/S2213-2600(18)30465-X

See **Articles** page 915

In *The Lancet Respiratory Medicine*, Nasser Altorki and colleagues[1] report a post-hoc analysis of perioperative morbidity and mortality experienced by participants in a randomised clinical trial (CALGB/Alliance 140503) designed to ascertain the optimum amount of lung tissue to be removed during surgical management of T1aN0 lung cancer. The primary endpoint of the study—disease-free survival—has not been reported yet because data are yet to mature, but at a minimum, this exploratory study provides an updated reference for expected adversity after surgery for early-stage lung cancer, albeit in the context of a clinical trial. Of 697 participants in the study, ten (1·4%) had died by 90 days after surgery, and adverse events of grade 3 or worse were reported in 102 (15%) patients. Altorki and colleagues point out that these results compare favourably with most published outcomes from large registries. To be fair, the study population probably benefited from several factors that have been associated previously with more favourable surgical outcomes—eg, healthier patients, skilled thoracic surgeons, and presumably high-volume study centres.[2,3] However, one could also argue that the surgical perspective offered by the clinical trial is, in some ways, more relevant than what has been used historically to characterise surgical safety. The available health-related information—eg, performance status, smoking history, and pulmonary function—suggests the patient population was similar to modern cohorts of surgically managed lung cancer patients. However, by requiring surgeons to comply with surgical standards in the CALGB/Alliance 140503 study (ie, preoperative workup, surgical lymph-node assessments, and surgical margin assessment), patients underwent oncological resections maintaining current best practices in thoracic oncology, which is less common in large registries.[4] Finally, the surgical environment—eg, surgeon training and hospital experience—is consistent with recommendations from medical societies and payers encouraging patients to have lung cancer surgery at high-volume hospitals, by surgeons who are skilled in thoracic surgery, using minimally invasive techniques.[5–7] As such, this cohort reflects the direction that surgically managed lung cancer is ideally heading.

The comparison of lobar (more lung tissue) and sublobar (less lung tissue) resections is highly relevant because the optimum extent of parenchymal treatment for early-stage lung cancer is currently unclear. Therefore, any additional risk associated with either of these two resection strategies would be of great clinical importance. Although the study by Altorki and colleagues is not the first to compare the safety of these different procedures, retrospective studies are subject to bias because less healthy patients are typically offered sublobar resections, potentially biasing against the sublobar cohort. By studying a randomised cohort of similarly healthy patients, Altorki and colleagues provide a novel perspective. Because the study was a post-hoc analysis, the infrequency of deaths renders the study considerably underpowered to exclude the observed 0·5% difference in 90-day mortality between lobar and sublobar resection. However, Altorki and colleagues assume that these two approaches would be judged similarly safe (by the oncology community and, presumably, patients). This