BJA
British Journal of Anaesthesia

Volume 125, Number 6, December 2020

British Journal of Anaesthesia, 125 (6): 843–862 (2020)
doi: 10.1016/j.bja.2020.07.040
Advance Access Publication Date: 21 August 2020

EDITORIALS

Bias and ethical considerations in machine learning and the automation of perioperative risk assessment

Vikas N. O'Reilly-Shah[1,2,*], Katherine R. Gentry[1,2], Andrew M. Walters[1], Joel Zivot[3], Corrie T. Anderson[1,2] and Patrick J. Tighe[4]

[1]Department of Anesthesiology and Pain Medicine, University of Washington School of Medicine, Seattle, WA, USA, [2]Department of Anesthesiology and Pain Medicine, Seattle Children's Hospital, Seattle, WA, USA, [3]Department of Anesthesiology, Emory University, Atlanta, GA, USA and [4]Department of Anesthesiology, University of Florida, Gainesville, FL, USA

*Corresponding author. E-mail: voreill@uw.edu

Keywords: artificial intelligence; gender bias; healthcare inequality; machine learning; perioperative medicine; racial bias; risk prediction

I know I've made some very poor decisions recently, but I can give you my complete assurance that my work will be back to normal. I've still got the greatest enthusiasm and confidence in the mission. And I want to help you.

HAL-9000 in 2001: A Space Odyssey

Anaesthesiologists have an ever-increasing number of tools at their disposal to assess the risk that a patient will assume when undergoing anaesthesia and surgery, and the likelihood of specific events during surgery. The earliest versions of these tools were based on heuristics or on a small number of easily identifiable features; the Revised Cardiac Risk Index (RCRI) and pulse pressure variation are good examples of this type of preoperative and intraoperative assessment tool. However, as medicine and technology have evolved, extremely large amounts of healthcare data have become available as computation and data storage capabilities have grown ever cheaper. This abundance of data and computational power has yielded the ability to calculate specific risk scores based on an increasing number of features, including risk calculators based on the National Surgical Quality Improvement Program (NSQIP)[1] and Society of Thoracic Surgeons (STS)[2] databases. In the context of the oft-quoted limitation on human cognitive capacity to 5–10 facts per decision, researchers have argued for the use of artificial intelligence (AI)/machine learning (ML) techniques to tap the full potential of this flood of data to achieve truly personalised medical decision support.[3]

Specifically in anaesthesiology, there is growth and great interest in the use of big data and ML for clinical care, particularly in the context of the coronavirus disease 2019 (COVID-19) pandemic.[4,5] However, there are a number of barriers and concerns not commonly discussed that anaesthesiologists should be aware of when encountering technologies that make use of AI for clinical decision support. Here we highlight a number of troubling aspects to the use of AI that are directly relevant to the practicing anaesthetist, specifically the propagation of racial and

gender bias and the relationship of risk modelling to informed consent. We close with some thoughts on how researchers, corporations, and regulators with interests in the use of AI for medical applications should address these issues through transparency in training data, mandatory disclosures, and appropriate labelling to allow adequate evaluation by clinicians.

Machine learning and propagation of bias

Risk scores, whether originating from traditional, linear statistical methods or non-linear methods such as deep learning neural networks, generally follow the same approach for classification tasks. First, a set of variables, or features, are collected for each patient in a large sample of patients. Each patient is also labelled with an outcome of interest: survival at 30 days after surgery, nausea on postoperative day 1, and so on. As a 'sample' (in the statistical sense of the term), this collection of data is presumed to be representative of a broader population of patients. The relationship between the patient features and their outcome is then linked using a statistical function so that for a given set of features that are added, subtracted, multiplied, or divided together (with variable levels of complexity in the underlying function), we can predict the risk of the outcome of interest. For traditional risk scoring using common statistical approaches, this may be a simple line or curve: understandable, predictable, but the relationship remains limited to a single, linear function. For deep learning, this may involve an incredibly complicated, non-linear hyperplane highly customised to the training set.

The first major problem is that data used for training models (ML or otherwise) can be readily biased. Mehrabi and colleagues[6] have identified 23 separate sources of bias including historical bias, representation bias, measurement bias, aggregation bias, content production bias, linking bias, sampling bias, and temporal bias. Additional biases can be introduced by the analytic approach; 'Simpson's paradox' is particularly instructive here.[7,8] The paradox occurs when a conclusion based on aggregated data is opposite to the conclusion when subgroups are analysed. It would be easy to imagine situations where a modelling approach that does not account for important confounders, variables associated with social determinants of health, for example, yields associations that are incorrect. Specifically with respect to risk score calculation, omitted variable bias presents another challenge. Busy clinicians may not have time to evaluate underlying model development processes, and therefore incorrectly assume that variables not included in a particular risk scoring scheme are non-contributory. In reality, the variable may not be included simply because it was never evaluated, having significance that is unknown rather than null. Biases in data are far from theoretical; in one example, more than 80% of subjects in a reference image dataset were light-skinned individuals, resulting in artificial vision algorithms that were unable to identify dark-skinned humans.[9] Another review showed how a facial recognition tool falsely classified 28 members of the US Congress as criminals, including 40% misclassification for members of colour despite only 20% representation.[10] Within the field of anaesthesiology, a preliminary multicentre analysis of data from 40 institutions by White and colleagues[11] revealed that Black patients received inferior care (with respect to postoperative nausea and vomiting prophylaxis) both in aggregate and individually at nearly every single centre. (NB: The choice to capitalize, or not, 'Black' race in an evolving standard and intentional choice meriting a discussion that is beyond the scope of this article).

Bias in ML is not restricted to just the data used for modelling. Many modern ML methods remain difficult to interpret; with hundreds of thousands of parameters used to generate a classification, determining what exactly leads to the predicted outcome becomes difficult for both theoretical and pragmatic reasons. The implications of model-based bias become more alarming with the advent of 'auto-machine learning' solutions which automate the process of model selection and tuning, potentially placing one black box within another. 'Fair' ML algorithms are those whose predictions remain independent of key features (e.g. sex, race, ethnicity) that, from a perspective of ethics, should not be associated with the outcome. It should be recognised that in some cases, such an association does exist and should be accounted for (e.g. age and sex in the prediction of postoperative pain or nausea). However, though some associations may result from physiological differences, other associations will result from differences in approach to treatment, including differences resulting from overt or implicit bias. These associations need to be detected and understood in order to account for actual risk in appropriate cases while calling out the presence of biased care in others.

Several solutions to mitigate bias have emerged. In the preprocessing stage, samples can be remapped to new representations in such a way as to remove information correlated to the sensitive attributes (e.g. race) but to preserve the overall relationship between the original data and the outcome of interest.[12] Model optimisation techniques to minimise bias include methods for regularisation or constraints on models at the time of model training. After model training, post-processing solutions to ML fairness such as the FairML library and 'What-if' tool can help identify key features which lead to the observed model predictions, allowing for human-level intervention in feature selection, model interpretation, and better understanding of model performance in different clinical scenarios.[13,14] These methods can also determine if the population-based model performs better in certain subgroups, an important step towards the personalisation of evidence-based perioperative management. Increasingly, data scientists recognise that ML models in production require continued attention. Underlying populations, decision, and outcome characteristics can change over time leading to concept drift, such that the original mappings of features to outcomes become less and less valid, and require updated training cycles to maintain accuracy and fairness.[15]

Ethics of risk prediction

A harmful chain of events can occur when algorithms that are inherently biased or founded on incorrect assumptions are used as the basis for clinical and system-wide decisions. This is the downstream impact of aforementioned biases in underlying data or modelling, where observed risk is caused by biased care in the underlying data rather than biologically plausible mechanisms for disparate risk. For instance, there are numerous clinical algorithms that include a race adjustment based on questionable or non-existent evidence.[16] 'Race-adjusted' scores can impact triage and treatment decisions that may exacerbate pre-existing health disparities and serve as self-fulfilling prophecies for people of colour. For instance, the American Heart Association's 'Get with the Guidelines' Heart Failure Risk Score assigns three additional points to patients who are 'non-Black' with no rationale provided.[17] If a White patient and a Black patient present with identical symptoms, the algorithm predicts the White patient to be at

higher risk of dying from heart failure, thereby encouraging physicians to allocate more resources to the White patient. A prediction tool used in patients with rectal cancer assigns Black patients a higher risk of cancer-related mortality than patients identifying as White or other.[18] This is based upon analysis of rectal cancer outcomes in a large sample of US patients, for which being Black was found to contribute to mortality risk. In this sample, an overwhelming majority of patients were White. Thus, the survival data for Black patients are already based upon a small, potentially biased sample. And further, we must ask whether it seems likely that being of African descent actually influences the biology of rectal cancer, or whether social determinants, such as access to healthcare, lifestyle factors, environmental factors, and mistrust of the healthcare system, might be more likely to underlie this increased risk of mortality for Black patients. Oncologists using this risk calculator, even those without overt or implicit bias against Black patients, may be less likely to recommend aggressive treatment to Black patients if they perceive that the patient's prognosis is poor. This pattern of interpreting risk through the lens of race, and subsequently limiting treatment, creates a self-fulfilling prophecy by which Black patients continue to have poorer health outcomes.

Incorrect assumptions and disparities in care can be perpetuated by algorithms that use surrogate measures of health status to predict future healthcare needs. In one example, past healthcare *spending* was used to predict which patients would benefit most from a program to coordinate their complex healthcare needs. However, investigators discovered that Black patients had a higher disease burden than Whites at the threshold at which enrolment in the program was recommended. Their analysis revealed that at a given level of disease burden, Blacks spent on average $1100–$1800 less annually than Whites.[19] This was interpreted by the algorithm as having a lower need for healthcare resources, when in fact Black patients had as much or a greater need for coordinated healthcare than their White counterparts.[19] There are numerous reasons that Black patients in the USA may not spend as much money on healthcare, ranging from having lower household income or being under- or uninsured, to facing practical barriers such as having to travel longer distances to access clinics and hospitals, having jobs with inflexible hours, and having childcare responsibilities.

A different ethical concern has to do with how physicians and healthcare systems utilise the information generated by risk prediction algorithms. One can imagine an ideal scenario in which a risk calculator facilitates shared decision-making and informed consent by generating patient-specific risk information. However, the converse could also occur: reliance on a risk calculator might curtail the practice of soliciting patients' values, prior experiences, and motivations when considering a future health intervention. And, physicians might limit the choices presented to patients/families when predicted risks are high, with paternalism trumping respect for autonomy. There is also a danger in health systems or insurance companies using risk calculators to determine which patients are appropriate candidates for certain interventions, circumventing shared decision-making by patients and doctors completely or penalising physicians who operate beyond the threshold of an 'acceptable' range of risk.

Finally, there is the question of liability for negligence. Though AI-based systems are US Food and Drug Administration-regulated medical devices that are designed to provide clinical decision support, not clinical decision-*making*, as these systems grow in complexity there is risk of increasing clinician reliance on them. In the event of a tort (harm) claim for negligence, a deviation of practice must result in a harm that is measured as some combination of loss of income and pain and suffering. In the USA, these tort claims require the opinion of a recognised expert according to the court by overcoming the Daubert test. That is, the evidence must be relevant and reliable. If harm results from a misapplication of an AI-derived scoring model, who is the defendant? Additionally, would an AI itself be able to provide evidence and pass a Daubert test?

Conclusions

Recognition of the impact of bias in AI modelling has spurred development of new tools allowing researchers to reexamine their datasets and potentially attenuate bias. Such tools can assist with data preparation to debias labelling[20] or word embeddings[21]; *post hoc* assessment of bias utilising newer methods such as counterfactual fairness; and re-weighing of data during training to mitigate algorithmic bias.[22] The work of Buolamwini and colleagues, including the work of the Algorithmic Justice League, deserves specific highlighting and commendation.[9] The US National Institute of Health's All of Us Research Program has been especially designed to address lack of diversity in underlying databases used in the development of healthcare AI and ML.[23] In addition to understanding the unique susceptibilities of various ML models to bias, researchers should be familiar with these tools, and their appropriate use should be considered at all stages of model development.

Publishers must also adapt to support a strong editorial and peer review process in journals that have long been dominated by RCTs and case control studies. Classic medical publication requirements such as baseline cohort patient characteristics are not sufficient to evaluate AI-based research. A more complete and transparent description of underlying datasets, and processes such as labelling and splitting, are vital to understand where sources of bias may emerge in a given model. To that end, guidelines such as CONSORT and TRIPOD are in the process of creating extensions specifically aimed at research utilising AI (CONSORT-AI and TRIPOD-ML); we expect rapid adoption of these guidelines by journals to improve quality control of published AI research.[24] The rapid proliferation of poorly constructed prediction models in the midst of the COVID-19 pandemic serves as a testament to the need for these interventions to be rapidly and widely adopted.[25]

Finally, providers consuming AI research need adequate understanding of its limitations and proper interpretation. Clinicians should be aware that research scientists and medical device companies have fallen short in these areas, and should maintain vigilance in their own practice to recognise the potential for bias in tools being used in their environments. The onus for this training will likely fall equally between training institutions for future providers and professional organisations for those in practice. Research curricula within teaching institutions which focus on basic statistics and biases in clinical research will need to develop new content on AI modelling, and how it compares with other forms of statistical analysis. Similarly, professional organisations should develop resources for providers to develop and test this knowledge through existing continuing education frameworks.

George Box famously wrote that all models are wrong, but some are useful. A corollary is that all models are wrong, but some are harmful. Returning to HAL-9000: its murderous actions were later revealed to have arisen from a multi-objective

optimisation problem with intractable tradeoffs. HAL was programmed to achieve three goals requiring mutually exclusive behaviour: to relay accurate information to the crew, to withhold the true nature of the mission from the crew, and to ensure mission completion. Similar themes underlie several examples of discrimination resulting from biased ML, where at-risk groups suffer disproportionately as a result of unexpected consequences, unanticipated tradeoffs, or both, all wrapped in the hubris of data science.

Authors' contributions

Meet **International Committee of Medical Journal Editors** (ICMJE) criteria for authorship: all authors Contributed to conception of the work and drafted and revised the manuscript: all authors.
Approved the final version to be published: all authors.

Declarations of interest

CTMA: part of Avanos Speaker's Bureau and a consultant for Fujifilm's ultrasound division. Shares in Abbvie and Masimo are held in joint brokerage account by CTMA and his wife. All other authors declare that they have no conflicts of interest.

Funding

United States National Institutes of Health NIH grants R01GM114290 (National Institute of General Medical Sciences), R01AG121647 (National Institute on Aging), and U24NS113800 (National Institute of Neurological Disorders and Stroke) (PJT).

References

1. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg* 2013; **217**: 833–42. e1–3
2. Fernandez FG, Shahian DM, Kormos R, et al. The society of thoracic surgeons national database 2019 annual report. *Ann Thorac Surg* 2019; **108**: 1625–32
3. Ahmed MN, Toor AS, O'Neil K, Friedland D. Cognitive computing and the future of health care cognitive computing and the future of healthcare: the cognitive power of IBM watson has the potential to transform global personalized medicine. *IEEE Pulse* 2017; **8**: 4–9
4. Char DS, Burgart A. Machine-learning implementation in clinical anesthesia: opportunities and challenges. *Anesth Analg* 2020; **130**: 1709–12
5. O'Reilly-Shah VN, Gentry KR, Van Cleve W, Kendale SM, Jabaley CS, Long DR. The COVID-19 pandemic highlights shortcomings in U.S. healthcare informatics infrastructure: a call to action. *Anesth Analg* 2020; **131**: 340–4
6. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. *A survey on bias and fairness in machine learning* 2019. Available from: http://arxiv.org/abs/1908.09635. [Accessed 13 July 2020]
7. Julious SA, Mullee MA. Confounding and Simpson's paradox. *BMJ* 1994; **309**: 1480–1
8. Kievit RA, Frankenhuis WE, Waldorp LJ, Borsboom D. Simpson's paradox in psychological science: a practical guide. *Front Psychol* 2013; **4**: 513
9. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editors. *Proceedings of the 1st conference on fairness, accountability and transparency*. New York, NY, USA: PMLR; 2018. p. 77–91
10. Nagpal S, Singh M, Singh R, Vatsa M. *Deep learning for face recognition: pride or prejudiced?* 2019. Available from: http://arxiv.org/abs/1904.01219. [Accessed 13 July 2020]
11. White R, Andreae MH, Ma X, et al. *Antiemetic prophylaxis as an anesthesia quality marker and its association with race in the multicenter perioperative outcomes group: a retrospective cohort study, 2004 - 2018*. Available from: http://www.asaabstracts.com/strands/asaabstracts/abstract.htm?year=2019&index=13&absnum=1500. [Accessed 15 July 2020]
12. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: *International conference on machine learning*; 2013. Atlanta, GA, USA
13. Adebayo JA. *FairML: ToolBox for diagnosing bias in predictive modeling*. Cambridge, MA, USA: Massachusetts Institute of Technology; 2016
14. Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viegas F, Wilson J. The what-if tool: interactive probing of machine learning models. *IEEE Trans Vis Comput Graph [Internet]* 2020; **26**: 56–65. Available from: https://doi.org/10.1109/TVCG.2019.2934619.
15. Benton WC. Machine learning systems and intelligent applications. *IEEE Softw* 2020; **37**: 43–9
16. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020. Available from: https://doi.org/10.1056/NEJMms2004740.
17. Peterson PN, Rumsfeld JS, Liang L, et al. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcome* 2010; **3**: 25–32
18. Bowles TL, Hu C-Y, You NY, Skibber JM, Rodriguez-Bigas MA, Chang GJ. An individualized conditional survival calculator for patients with rectal cancer. *Dis Colon Rectum* 2013; **56**: 551–9
19. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366**: 447–53
20. Jiang H, Nachum O. Identifying and correcting label bias in machine learning. *arXiv [cs.LG]* 2019. Available from: http://arxiv.org/abs/1901.04966. [Accessed 14 July 2020]
21. Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. *Advances in neural information processing systems 29*. Curran Associates, Inc.; 2016. p. 4349–57
22. Amini A, Soleimany AP, Schwarting W, Bhatia SN, Rus D. Uncovering and mitigating algorithmic bias through learned latent structure. In: *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*. *New York, NY, USA*; 2019. p. 289–95
23. Denny JC, Devaney SA, Gebo KA. The 'All of Us' research program. *Reply*. *N Engl J Med [Internet]* 2019; **381**: 1884–5. Available from: https://doi.org/10.1056/NEJMc1912496.
24. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019; **25**: 1467–8
25. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369**: m1328