Clinical applications of machine learning algorithms: beyond the black box

To maximise the clinical benefits of machine learning algorithms, we need to rethink our approach to explanation.

David S. Watson[1,2,3], Jenny Krutzinna[1], Ian N. Bruce[4,5], Chistopher E.M. Griffiths[5,6], Iain B. McInnes[7], Michael R. Barnes[2,3], Luciano Floridi[1,3]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK
[2]Centre for Translational Bioinformatics, William Harvey Research Institute, Queen Mary University of London, London, UK
[3]The Alan Turing Institute, London, UK
[4]Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology Medicine and Health, The University of Manchester, Manchester, UK
[5]NIHR Manchester Biomedical Research Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester M13 9WU, UK
[6]The Dermatology Centre, Salford Royal NHS Foundation Trust, The University of Manchester, Salford, UK
[7]Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, UK
Correspondence to: D Watson david.watson@oii.ox.ac.uk

Preprint not peer reviewed

Introduction

Machine learning algorithms are an application of artificial intelligence designed to automatically detect patterns in data without being explicitly programmed. They promise to change the way we detect and treat disease and will likely have a major impact on clinical decision making. The long term success of these powerful new methods hinges on the ability of both patients and doctors to understand and explain their predictions, especially in complicated cases with major healthcare consequences. This will promote greater trust in computational techniques and ensure informed consent to algorithmically designed treatment plans.

Unfortunately, many popular machine learning algorithms are essentially black boxes—oracular inference engines that render verdicts without any accompanying justification. This problem has become especially pressing with passage of the European Union's latest General Data Protection Regulation (GDPR), which some scholars argue provides citizens with a "right to explanation." Now, any institution engaged in algorithmic decision making is legally required to justify those decisions to any person whose data they hold on request, a challenge that most are ill equipped to meet. We urge clinicians to link with patients, data scientists, and policy makers to ensure the successful clinical implementation of machine learning. We outline important goals and limitations that we hope will inform future research.

Predictions versus explanations

Predictions tell us that x is true; explanations tell us why x is true. The past decade has seen enormous advances in our ability to predict complex phenomena using computational techniques. Explanatory breakthroughs, on the other hand, have been few and far between.

Machine learning algorithms have already shown expert diagnostic performance based on imaging data for conditions including diabetic retinopathy,[1] skin cancer,[2] and pneumonia.[3] Precision medicine seeks to go further, modelling molecular data to classify patients according to endotype,[4] defining disease mechanism and ontologies.[5] With the integration of electronic health records and wearable medical sensors, machine learning may usher in a new era of real time diagnostic updates, enabling earlier, more targeted interventions.[6]

Preprint not peer reviewed

Machine learning techniques are already emerging in clinical practice.[7] Microsoft's InnerEye offers a graphical user interface to algorithms that help radiologists diagnose cancerous tumours and plan precise surgical interventions.[8] DeepMind Health recently partnered with Moorfields Eye Hospital to develop models for diagnosing common retinal pathologies based on optical coherence tomography scans.[9] IBM's Watson for Oncology seeks to provide personalised cancer care based on health records,[10] although the project has run into numerous procurement problems, cost overruns, and delays.[11]

One frequently cited obstacle to machine learning's wider clinical adoption is a lack of understanding among patients and doctors about how predictions are made.[12] This is especially true of some top performing algorithms, like the deep neural networks used in image recognition software. These models may reliably discriminate between malignant and benign tumours, but they offer no explanation for their judgments. Of course, clinicians are not always able to perfectly account for their own inferences, which may be based more on experience and intuition than explicit medical criteria.[13] Moreover, doctors do not optimally integrate all available evidence, and cognitive biases can be deeply entrenched.[14] Still, many think that, as a new technology, the burden of proof is on machine learning to account for its predictions. If doctors do not understand why the algorithm made a diagnosis, then why should patients trust the recommended course of treatment? Is informed consent even possible without some grasp of how the model reached its conclusion?

Not all algorithms are black boxes. Some sophisticated models, such as those based on regularised linear regression, provide a modest number of informative parameters.[15] Yet, although restricting the use of clinical machine learning to more intelligible algorithms is tempting, it would be a mistake. No single technique is optimal for all cases—a result known as the "no free lunch theorem" in computer science[16]—which means that any attempt to shoehorn all datasets into a particular family of statistical models is guaranteed to fail.

The opportunity costs of not using our best available tools for disease detection and treatment are substantial—12 million people a year receive misdiagnoses in the United States, with about six million facing potential harm as a result.[17] Nearly one third of all preventable deaths in the United Kingdom are attributable to misdiagnosis.[18] The benefits of early disease detection are well known.

Preprint not peer reviewed

Yet clinicians are right to be sceptical of inscrutable models. Especially worrisome is the risk of overfitting to an unrepresentative sample. In one famous example, an algorithm designed to predict probability of death among hospital patients with pneumonia systematically classified asthmatics as low risk.[19] The correlation was spurious—patients with asthmatic pneumonia were sent directly to the intensive care unit, where they received continuous treatment that improved their prognosis so substantially that they seemed to have better than average chances of survival. Mistakes like this show the potential dangers of naively accepting the outputs of a black box model. They also raise important questions about liability in cases of algorithmic error. Who is ultimately responsible for a computational misdiagnosis? Clinicians? Data scientists? Policy makers have tackled similar questions in other contexts and come to no clear consensus.[20]

Right to explanation?

The GDPR has emphasised "explainability" as a top priority in machine learning research, provoking a global debate over the right to explanation in cases where individuals are subject to automated decisions. Whether or not this purported right is enshrined in the GDPR—a point of contention among legal scholars[21]—there are compelling reasons to endorse it in medical contexts. This will require a total reorientation of priorities for data scientists, who are more used to optimising for accuracy than for intelligibility.

Before we can design new methods to tackle this challenge, we must agree on what constitutes a satisfactory explanation. Do we want to understand all the patterns the machine has learnt (model-centric explanations)? Or just those that are relevant to the patient (subject-centric explanations)?[22] The former aims to provide global understanding about the relative importance of all variables and how they interact to make predictions, which may shed new light on disease mechanisms; the latter provides local understanding about why this particular input led to that particular output, which could be relevant for individual patient prognosis.

Clinicians sceptical of machine learning tend to focus on the lack of clear model-centric explanations.[19] Deep neural networks, for example, routinely contain millions of parameters, assigning weights and biases to thousands of nodes in an architecture so

Preprint not peer reviewed

complex that no human could plausibly be expected to grasp the whole model's internal mechanics. But if a computer truly outperforms doctors in making diagnoses, then we would like to know why. Understanding the biological patterns or processes it has uncovered could advance our knowledge and help build the medical community's trust in such systems.

Of course, patients are the most critical stakeholders in clinical machine learning. Enabling them to appreciate their algorithmically determined diagnosis and treatment options is crucial—but also complicated, especially when inputs include high dimensional genomics or imaging data. Researchers in the nascent field of interpretable machine learning have implemented methods for generating model-agnostic local explanations.[23, 24] These approaches are promising, but more work is needed to extend them to clinical settings and support them with the appropriate medical ethics framework.

The path forward

Current proposals struggle to meet two important criteria for the clinical application of machine learning: scalability and customisability. With biological datasets often containing millions of variables per sample, the computational complexity of explanatory methods for molecular models must be constrained. This entails an inherent trade-off between completeness and simplicity. Ideally, users could specify a level of explanatory granularity that best suits their needs. Some may prefer diagnoses to be explained in terms of basic, familiar biological concepts; others may opt for a more detailed account in terms of molecular mechanisms.

Important unanswered questions remain about how best to measure the utility of a given explanation. Some authors have attempted to formalise the problem in a computable fashion,[25] whereas others advocate a more empirical approach driven by experimental psychology.[26] Both methods have their merits and drawbacks, but building a research programme around either will be difficult without first establishing a broad consensus.

Some caution with regard to transparency is advisable. A fully open source approach may enable misuse of the algorithm for harmful purposes outside the clinical context. This is particularly problematic when a diagnosis is based on easily accessible data, such

Preprint not peer reviewed

as facial images or movement patterns.[27] Scrutiny of machine learning is important but should not expose people to disproportionate risks or privacy violations, especially when there is no immediate benefit to diagnosis, as is the case with currently untreatable conditions.

We are only just beginning to realise machine learning's potential for medicine, and although it remains exploratory the benefits should not be ignored. Patients, clinicians, and data scientists must collaborate to develop new methods for extracting model-centric and patient-centric explanations that can provide global and local understanding. Bringing algorithms into the clinic can advance knowledge and improve care, but only if we are prepared to devote sufficient resources to illuminating the black box for doctors and patients alike.

Preprint not peer reviewed

References

1. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 2016; 316: 2402-10. PubMed doi:10.1001/jama.2016.17216.
2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-8. PubMed doi:10.1038/nature21056.
3. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-level pneumonia detection on chest X-Rays with deep learning. *arXiv* 2017; 1711.05225. https://arxiv.org/abs/1711.05225.
4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2014; 13: 8-17. PubMed doi:10.1016/j.csbj.2014.11.005.
5. Kibbe WA, Arze C, Felix V, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015; 43(D1):D1071-8. PubMed doi:10.1093/nar/gku1011.
6. Obermeyer Z, Emanuel EJ. Predicting the future–big data, machine learning, and clinical medicine. *N Engl J Med* 2016; 375: 1216-9. PubMed doi:10.1056/NEJMp1606181.
7. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis* 2018; 66: 149-53. PubMed doi:10.1093/cid/cix731.
8. Sample I. "It's going to create a revolution": how AI is transforming the NHS. *Guardian* 2018 https://www.theguardian.com/technology/2018/jul/04/its-going-create-revolution-how-ai-transforming-nhs.
9. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018; 24: 1342-50. PubMed doi:10.1038/s41591-018-01076.
10. Somashekhar SP, Sepúlveda MJ, Publielli S, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol* 2018; 29: 418-23. PubMed doi:10.1093/annonc/mdx781.
11. Schmidt C. M.D. Anderson Breaks with IBM Watson, Raising Questions about Artificial Intelligence in Oncology. *J Natl Cancer Inst* 2017; 109:djx113. PubMed doi:10.1093/jnci/djx113.

Preprint not peer reviewed

12. Lipton Z. The doctor just won't accept that! *arXiv* 2015; 1711.08037v2. https://arxiv.org/abs/1711.08037.
13. Mukherjee, S. AI versus MD. *The New Yorker* 2017 https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.
14. Gigerenzer G, Gaissmaier W. Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest*. 2007; 8:53-96.
15. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc. Series B (Methodological)*. 1996; 58: 267-88. doi:10.1111/j.2517-6161.1996.tb02080.x.
16. Wolbert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997; 1: 67-82. doi:10.1109/4235.585893.
17. Singh H, Meyer A, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual Saf* 2014; 23: 727-31. PubMed doi: 10.1136/bmjqs-2013-002627.
18. Hogan H, Healey F, Neale G, Thomson R, Vinent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Qual Saf* 2012; 21: 737-45. PubMed doi:10.1136/bmjqs-2011-001159.
19. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2015; 1721-30.
20. Johnson DG, Verdicchio M. Ai agency and responsibility: the VW fraud case and beyond. *AI Soc* 2018, doi:10.1007/s00146-017-0781-9.
21. Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 2017; 7: 76-99. doi:10.1093/idpl/ipx005.
22. Edwards L, Veale M. Slave to the algorithm? Why a right to explanation is probably not the remedy you are looking for. *Duke Law and Technology Review* 2017; 16: 18-84.
23. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *arXiv* 2016; 1602.04938v3. https://arxiv.org/bas/1602.04938.
24. Lundberg S, Lee S. A unified approach to interpreting model predictions. *arXiv* 2017; 1705.07874v2. https://arxiv.org/abs/1705.07874.
25. Lakkaraju H, Kamar E, Caruana R, Leskovec J. Interpretable and explorable approximations of black box models. *arXiv* 2017; 1707.01154. https://arxiv.org/abs/1707.01154.

Preprint not peer reviewed

26. Doshi-Velez F, Been K. Towards a rigorous science of interpretable machine learning. *arXiv* 2017; 1702.08608v2. https://arxiv.org/abs/1702.08608.

27. Hallowell N, Parker M, Nellåker C. Big data phenotyping in rare diseases: some ethical issues. *Genet Med* 2019; 21: 272-4. PubMed doi:101038/s41436-018/0067-8.

Preprint not peer reviewed