Reality Check: The Limitations of Artificial Intelligence in Clinical Medicine

BENJAMIN JONES
MATT ARCHER
STEPHANIE GERMAIN

Author affiliations can be found in the back matter of this article

EDITORIAL

IJS Press
Part of the IJS Publishing Group

ABSTRACT

Artificial intelligence is poised to transform clinical medicine, yet for successful implementation to occur we must also appreciate its limitations. The heterogeneity of current research, particularly in relation to the use of data, means that results cannot necessarily be extrapolated to a population level. Robust study designs are required to minimise the introduction of bias into artificial intelligence models and generate a strong body of evidence. Identifying the specific areas of healthcare where artificial intelligence can have the greatest impact will be essential in ensuring it has a positive influence on clinical outcomes and patient experience over the coming years.

CORRESPONDING AUTHOR:
Benjamin Jones
Royal Berkshire Hospital, GB
Ben_jones95@live.co.uk

KEYWORDS:
artificial intelligence; machine learning; data; limitations; clinical implementation

TO CITE THIS ARTICLE:
Jones B, Archer M, Germain S. Reality Check: The Limitations of Artificial Intelligence in Clinical Medicine. *International Journal of Digital Health*. 2021; 1(1): 8, 1–4. DOI: *https://doi.org/10.29337/ijdh.32*

The rise of digitalisation has undoubtedly transformed the way that we live and work. Vast quantities of data are produced every minute, including personal correspondence, internet searches and electronic patient records. The generation of large volumes of complex information renders traditional methods of data interpretation and visualisation insufficient for detailed analysis. The field of big data has emerged to address these challenges. At the forefront is artificial intelligence (AI), a term used to describe devices able to respond to their environment and select actions to achieve particular goals.

The primary focus of AI research at present is machine learning (ML), a field in which algorithms create their own mathematical model based upon sample data through recursive self-improvement. Within ML, a vast array of approaches exist, often differentiated by the model architecture or learning paradigm involved (***Table 1***). The clinical application of ML has already produced astounding results, in radiology for instance, where AI has been demonstrated to detect breast cancer from mammograms more accurately than radiologists [1]. A paradigm shift appears to be on the horizon.

If we are to incorporate AI into clinical practice effectively, we must also evaluate its limitations. Ethical dilemmas, implementation challenges and adoption-related risks are all key considerations. In this article, we focus on technical limitations and how these may shape the role of AI in healthcare. The subsequent paragraphs address issues surrounding the quality and quantity of data used in ML, problems with the introduction of greater numbers of variables, and approaches to clinical implementation that can help to mitigate the drawbacks of AI.

AI performance is largely dependent on the amount of available data that can be fed into the model. Whilst a larger dataset provides more information from which a machine can learn patterns, more data is not necessarily better. For example, class imbalance is a common problem whereby a trained model favours the prediction of the majority class, demonstrated by algorithms trained to predict driver mutations in glioblastoma and ovarian carcinoma [5]. This occurs because trained models can achieve higher accuracy by favouring the majority class yet perform poorly at detecting rarer classes which are often those of greater clinical significance in the first place [6].

The paucity of any evidence-based guidance on the size of datasets represents another major hurdle we must overcome if we are to standardise ML in clinical medicine [7]. As sample size can vary drastically between different studies, it would be pertinent to introduce a framework through which optimum sample size for a given study can be determined. Furthermore, many publications use data from a single hospital, region, or country, resulting in models that cannot produce repeatable prediction accuracy when extrapolated across larger or more diverse datasets [8]. This phenomenon is often referred to as "overfitting" and occurs when AI places undue significance on indiscriminate features, introducing a source of bias into the model.

Even if we can curate high-quality datasets, accumulating historical data may not actually be of benefit in many scenarios. Indeed, evidence has demonstrated that the accuracy of predicting future events in clinical care may actually be worse if we include more longitudinal data, with recent data alone being the best barometer [9]. Such predictions in clinical medicine are akin to financial models that have failed to predict future changes in the stock market; historical patterns are often redundant as they generally have little bearing on the likelihood of future events.

These issues highlight the relative value of ML in diagnostics and image analysis, where the outcome we are training the model to predict is unlikely to change significantly in the future, for instance, the appearance of a lung cancer on a computerised tomography (CT) scan. Developing complex tools that are able to predict future health events or select the optimum management approach for a specific patient are much more challenging. Not only do more factors need to be taken into account when solving more nuanced problems, but

LEARNING PARADIGM	DESCRIPTION	CLINICAL APPLICATIONS
Supervised	Algorithm uses training data consisting of input features matched with a target output to create a model able to predict the output value of new input data based on the relationships it has learned.	Pattern recognition, for example, predicting mortality through risk stratification in COVID-19 patients [2].
Unsupervised	Algorithms identify relationships within a dataset for which there are no pre-existing categories or target output labels.	Discovering new patterns in clinical data, for example, identifying features that may contribute to the progression of Alzheimer's disease [3].
Reinforcement	Iterative learning process in which the algorithm continuously observes outcomes from a particular action within an environment and selects future actions in order to maximise reward.	Automatisation of complex tasks, for instance, controlling the infusion rate of a general anaesthetic during surgery through the monitoring of physiological measurements [4].

Table 1 Learning paradigms used in ML and their applications within clinical medicine.

we are also chasing a moving target by virtue of the fact we exist in a dynamic world with an evolving healthcare landscape. Sceptics might argue that if AI is not able to predict future trends, merely recapitulate historical patterns, then it is in fact not "intelligent" at all.

A potential solution to the inability of AI to predict complex phenomena is the addition of more variables, ranging from genomics to demographics, and even lifestyle factors such as exposure to ultra-violet light or consumption of red wine. However, the incorporation of vast detail may give rise to a "butterfly effect". Popularised by Edward Lorenz in the 1960s in reference to the modelling of weather systems, it describes the phenomenon in which small differences in the initial state of a system can lead to large differences in a later state. He found that minute rounding errors had drastic effects on future weather patterns, leading him to postulate that the movement of air caused by the wings of a butterfly could result in a tornado. By exposing AI to a greater number of variables, we are generating models that are more prone to overfitting [10], limiting their utility in generating robust predictions that can be generalised at a population level.

It is possible to mitigate the shortfalls of AI to an extent by stratifying complex outcomes into discrete variables. Instead of asking "When will this patient have a myocardial infarction?" we could reframe the question by training an AI to predict the annual risk of a cardiovascular event. Such risk stratification tools are already widespread throughout clinical practice and ML has the potential to enhance these. Indeed, ML models have been shown to be superior to traditional statistical models in the prediction of major cardiovascular events, heart failure and cardiac dysrhythmias [11].

Nonetheless, for AI to be integrated into clinical medicine, a body of evidence matching current standards of scientific rigour is required. There is a scarcity of prospective studies and randomised-controlled trials evaluating ML, even in medical imaging, where there is considerable optimism that implementation may be on the horizon [12]. Prospective randomised clinical trials are essential to mitigate the various sources of bias which have the potential to confound results. Given the potential of ML to exaggerate bias through overfitting, retrospective evidence is simply not good enough if we are to utilise AI safely in clinical practice. Furthermore, as model complexity increases, the data relationships learned become more elaborate and harder to extract. This so-called "black-box problem" has enormous ramifications for medicine, as we cannot justify decisions made by AI if we are uncertain whether hidden bias may exist within the algorithm.

AI is poised to transform clinical medicine, yet it is crucial we appreciate current limitations. For the foreseeable future, the role of ML will likely remain limited to pattern recognition and task automation, where the vastly superior processing power of machines provides an enormous advantage over the human brain. The successful implementation of AI in clinical practice will require issues surrounding data and bias to be addressed. A bright future may lie ahead, but it is vital we remember AI is not yet a panacea.

COMPETING INTERESTS

BJ has completed paid consultancy work for ni2o. All other authors have no competing interests.

AUTHOR AFFILIATIONS

Benjamin Jones *orcid.org/0000-0003-0257-6049*
Royal Berkshire Hospital, GB

Matt Archer *orcid.org/0000-0001-8035-2783*
Royal Hampshire Hospital, GB

Stephanie Germain
Kent Surrey and Sussex Deanery: Health Education England Kent Surrey and Sussex, GB

REFERENCES

1. **McKinney SM,** *et al.* 'International evaluation of an AI system for breast cancer screening'. *Nature*, 577(7788): Art. no. 7788, Jan. 2020. DOI: *https://doi.org/10.1038/s41586-019-1799-6*
2. **Gao Y,** *et al.* 'Machine learning based early warning system enables accurate mortality risk prediction for COVID-19'. *Nat. Commun.*, 11(1): Art. no. 1, Oct. 2020. DOI: *https://doi.org/10.1038/s41467-020-18684-2*
3. **Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA.** 'The Application of Unsupervised Clustering Methods to Alzheimer's Disease'. *Front. Comput. Neurosci.*, 13: 31, 2019. DOI: *https://doi.org/10.3389/fncom.2019.00031*
4. **Meskin N, Haddad W, Padmanabhan R.** 'Closed-Loop Control of Anesthesia and Mean Arterial Pressure using Reinforcement Learning'. *Biomed. Signal Process. Control*, 22: 54–64, Sep. 2015. DOI: *https://doi.org/10.1016/j.bspc.2015.05.013*
5. **Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K.** 'CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features'. *PLOS ONE*, 8(10): e77945, Oct. 2013. DOI: *https://doi.org/10.1371/journal.pone.0077945*
6. **Valdebenito J, Medina F.** 'Machine learning approaches to study glioblastoma: A review of the last decade of applications'. *Cancer Rep.*, 2(6): e1226, 2019. DOI: *https://doi.org/10.1002/cnr2.1226*
7. **Balki I,** *et al.* 'Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A

Systematic Review'. *Can. Assoc. Radiol. J. J. Assoc. Can. Radiol.*, 70(4): 344–353, Nov. 2019. DOI: *https://doi.org/10.1016/j.carj.2019.06.002*

8. **Lanka P, Rangaprakash D, Dretsch MN, Katz JS, Denney TS, Deshpande G.** 'Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets'. *Brain Imaging Behav.*, 14(6): 2378–2416, Dec. 2020. DOI: *https://doi.org/10.1007/s11682-019-00191-8*
9. **Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB.** 'Decaying Relevance of Clinical Data Towards Future Decisions in Data-Driven Inpatient Clinical Order Sets'. *Int. J. Med. Inf.*, 102: 71–79, Jun. 2017. DOI: *https://doi.org/10.1016/j.ijmedinf.2017.03.006*
10. **Nichols JA, Herbert Chan HW, Baker MAB.** 'Machine learning: applications of artificial intelligence to imaging and diagnosis'. *Biophys. Rev.*, 11(1): 111–118, Feb. 2019. DOI: *https://doi.org/10.1007/s12551-018-0449-9*
11. **Patel B, Sengupta P.** 'Machine learning for predicting cardiac events: what does the future hold?' *Expert Rev. Cardiovasc. Ther.*, 18(2): 77–84, Feb. 2020. DOI: *https://doi.org/10.1080/14779072.2020.1732208*
12. **Nagendran M,** *et al.* 'Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies'. *BMJ*, 368: m689, Mar. 2020. DOI: *https://doi.org/10.1136/bmj.m689*

TO CITE THIS ARTICLE:
Jones B, Archer M, Germain S. Reality Check: The Limitations of Artificial Intelligence in Clinical Medicine. *International Journal of Digital Health*. 2021; 1(1): 8, 1–4. DOI: *https://doi.org/10.29337/ijdh.32*

Submitted: 29 January 2021 Accepted: 21 March 2021 Published: 12 April 2021