DIGITAL TECHNOLOGY

# Key considerations for the use of artificial intelligence in healthcare and clinical research

**Authors:** Christopher A Lovejoy,[A] Anmol Arora,[B] Varun Buch[C] and Ittai Dayan[D]

ABSTRACT

**Interest in artificial intelligence (AI) has grown exponentially in recent years, attracting sensational headlines and speculation. While there is considerable potential for AI to augment clinical practice, there remain numerous practical implications that must be considered when exploring AI solutions. These range from ethical concerns about algorithmic bias to legislative concerns in an uncertain regulatory environment. In the absence of established protocols and examples of best practice, there is a growing need for clear guidance both for innovators and early adopters. Broadly, there are three stages to the innovation process: invention, development and implementation. In this paper, we present key considerations for innovators at each stage and offer suggestions along the AI development pipeline, from bench to bedside.**

**KEYWORDS:** machine learning, innovation, technology, algorithms, data

**DOI:** 10.7861/fhj.2021-0128

## Introduction

Despite considerable acclaim, promotion and investment around artificial intelligence (AI), the technology is simply a form of computational analysis not sharply demarcated from other kinds of modelling.[1–3] AI is defined as the ability of a computer system to perform tasks that are usually thought to require human intelligence, including processes of learning, reasoning and self-correction.[4] The term is sometimes used synonymously with 'machine learning', a methodology that allows a computer system to refine its output function by learning from input data, with minimal human intervention. The strength of AI is its ability to accentuate the flexibility and expressivity of more traditional statistical techniques, catering to problems where the inputs and outputs are highly multi-dimensional and associated in extremely complex ways. A number of guidelines currently exist for the development of AI solutions, including regarding concerns about transparency, reproducibility, ethics and effectiveness.[5,6] Reporting guidelines have also been produced in order to assess the methodology of projects.[7,8] However, real-world requirements for AI solutions are still evolving as noted in a recent UK parliamentary research briefing on the topic of AI in healthcare.[9] In the absence of any set precedent for its clinical application, it is essential that researchers have an appreciation of AI's strengths and limitations to assist them in developing appropriate research questions. It is well established that innovation consists of three stages: invention, development and implementation.[10] Here, we present key considerations at each of the major stages in the AI innovation pathway, from identifying a research question to deployment.

## Identifying appropriate research questions

Before embarking on AI-based research, a fundamental and often overlooked question is: 'Would AI really be appropriate for the research question at hand?' There are research projects that can be enhanced with AI and others where AI can be detrimental. AI is useful in situations where the input is highly complex, where it is desirable for the question to have a complex answer or where the hypothesis space cannot easily be constrained. An example of a highly complex input is imaging data, which may contain numerous features that can represent a large number of pathologies.

The range of uses of AI varies from automated data collection to developing diagnostic decision aids. The techniques required for each are similar and revolve around the ability of models to identify complex relationships within datasets due to their capacity to analyse many variables and their ability to extract useful features; for example, with minimal human instruction, convolutional neural networks (CNNs) can extract features from medical images, such as recent work with histology slides.[11,12] CNNs, which are modelled on the animal visual cortex, are particularly suited to analysis of imaging data that may otherwise be difficult to analyse.

Rather than population-level scoring systems that provide generalisable but imprecise predictions for the non-existent 'average' patient, AI models can provide predictions more specific to smaller patient cohorts and with greater precision. Several models have illustrated this in recent months. Hilton *et al*'s model provides personalised predictions of adverse events, such as extended length of stay, 30-day readmission and death, while other recent studies provide personalised predictions of heart failure outcomes and risks of gestational diabetes or myocardial infarction.[13–16]

**Authors:** [A]physician, University College London, London, UK and University College Hospital, London, UK; [B]medical student, University of Cambridge, Cambridge, UK and honorary research fellow, Moorfields Eye Hospital, London UK; [C]director of AI development, MGH & BWH Center for Clinical Data Science, Boston, USA; [D]lecturer, Partners HealthCare, Boston, USA and chief executive officer, Rhino Health, Boston, USA

As well as being used to guide patient care, AI may be used to expand the scope of research by enabling automation of routine tasks; for example, if a team wished to explore the prevalence of pulmonary nodules, the time and financial costs required to manually annotate a large dataset of computed tomography (CT) could exceed the resources available to a small research group. However, using AI, if enough data were already accurately labelled, the researchers could train a classifier that could be used to analyse the remaining images. Similarly, AI may be applied to interpret other investigations or even to analyse clinic letters using natural language processing tools.

### Unhelpful AI

Not all problems need an AI-based solution. A common pitfall in industry is to search for solutions which utilise AI rather than focusing on existing problems. Such an approach is ill-advised because, aside from questionable clinical utility of the outputs, AI-based research has a number of inherent disadvantages. Ethical issues (such as algorithmic bias, lack of transparency and ambiguous accountability) have gathered attention as potential barriers to real-world adoption of AI systems.[17]

Firstly, large datasets often lack diversity and studies based on these datasets may not reflect the target population; for example, the UK Biobank excludes young people and has low numbers of several common diseases of interest, such as stroke.[18] While the algorithms may demonstrate superior performance on a limited set of test data, they may underperform when subjected to external validation on unseen data. This algorithmic bias may be seen with any predictive model, but AI models are particularly vulnerable as they may discriminate against certain patient groups while still maintaining very high aggregate performance measures, such as accuracy and area under receiver operating characteristic curve (AUC).

Secondly, common AI models, such as deep neural networks, have internal logic which is inherently difficult to interpret. This 'black-box' problem makes models more difficult to explain to patients, to interrogate when clinical intuition contradicts them and to improve in a systematic and rigorous manner.[19] For this reason, AI attracts greater regulatory scrutiny, which can present additional hurdles and uncertainty compared with conventional solutions. There is, however, active research into the development of methods to produce AI models while avoiding the 'black-box' phenomenon, including using local interpretable model-agnostic explanations (LIME).[20]

Thirdly, there are times when clinical decisions are entrusted to healthcare professionals and use of AI may be inappropriate; examples include decisions relating to withdrawing life-supporting treatment, decisions involving particularly sensitive clinical data (such as sexual history or infection status) and decisions where there is a risk of discriminatory bias.

## Engineering the model

### Collection and preparation of data

#### *How much data is needed?*

While efforts have been made to predict dataset size requirements, the precise amount required for a particular task is an inexact science and varies depending on the number of variables and the outcomes being studied.[21] In general, increasing the size of a training dataset increases performance, albeit with diminishing returns, whereby each incremental unit rise in dataset size produces a smaller improvement in performance. An empirical approach is generally recommended, increasing the dataset size until satisfactory clinical performance is achieved.

#### *How to obtain the dataset?*

An important guiding principle is to reduce bias and improve generalisability by sampling across the entire domain of intended use. Having data sourced from a wide-ranging demographic of patients, multiple geographical sites and with a large variety of presentations is always preferable, but not always possible. Collecting data for validation prospectively is also preferable to enable a greater chance of a robust and unbiased validation for the model.

Patient consent should be obtained and the risks and benefits of participation in research should be clearly explained. Notably, there are emerging risks to patient privacy from AI systems being developed that are capable of deanonymising patient data. It has already been possible to identify individuals from their electroencephalography and it has been suggested that clinical data (such as fundoscopy or electrocardiography) may contain more hidden information that humans are able to interpret.[22–24] Anonymising patient data is an increasingly important and necessary task and there are emerging methods to assist with this, including generating noise in the data using generative adversarial networks.

Consultation with a data scientist prior to collection is highly recommended because there may be nuances in the data labelling strategy or methods for optimising data collection that can be valuable to know before data collection commences.

#### *Defining the ground truth*

A ground truth is the answer to the question the model is being asked for each datapoint in the training dataset; for example, if the model is tasked with classifying skin lesions as cancerous or non-cancerous, the ground truth is the label assigned to each image. If the ground truth was established using only the labelling of an individual human, the maximal performance of the AI is limited to the accuracy of that human's labelling. An alternative method involves labelling the data using more parameters than the algorithm is being trained with or by consensus opinion of a committee of experts. A practical example may include training an algorithm to visually classify skin lesions but labelling the data based on more robust biopsy results rather than human visual classification. Other methods of enhancing safety of AI models include human-in-the-loop learning, which involves a human in the training, tuning and testing of an algorithm rather than relying upon a fully automated system.

### The multidisciplinary team

These projects are a multidisciplinary effort, in which two components are essential: clinical expertise and machine learning (ML) scientists. Clinicians are needed to help frame the clinical question, collect and annotate the data. ML expertise is needed for development and assessment of the model.

ML expertise may be provided by scientists within local hospitals, associated research institutions or through collaboration with commercial organisations. Data sharing and privacy considerations are particularly important when data are being shared beyond the boundaries of the host institution. The significant computing power required for training ML models may require collaboration with third party cloud-based computer systems. Interestingly, it has been estimated that the computing power required to train large AI models doubled every 3.4 months between 2012 and 2018, with computing power being suggested as a potential roadblock to future AI development.[25,26]

## Deploying the model

### Generalisability

A model is only as useful as its ability to perform on novel data. While one strength of AI models is their ability to fit to complex nuances within a dataset, this places them at greater risk of finding artefactual idiosyncrasies in the dataset that are not reflective of a wider reality. Overfitting to these artefacts affects the model's ability to maintain useful performance when applied to unseen data. To mitigate the risk of overfitting and algorithmic bias, there must be sufficient diversity within the training dataset. The training data must be at least as diverse as the population that the algorithm intends to serve. External validation on independently derived data is required in order to ensure that the systems perform effectively when exposed to novel data.

### Regulation

AI models that are used as an integral part of the diagnosis and management of human disease are treated by the US Food and Drug Administration (FDA) as 'Software as a Medical Device' (SaMD). The regulatory standards for SaMD are still evolving, but a few principles are becoming apparent. Unlike traditional medical devices, which would not change after development, AI algorithms can be updated and improved as new data are collected. Good performance at the time of deployment does not guarantee that the model will continue to perform well. This introduces the need to regulate throughout the lifetime of an algorithm, and the need to continually demonstrate safe and effective practices. At present, it is unclear when and how often the FDA will review these algorithms but embedding high-quality engineering practices, such as data traceability and regular performance review, will help to demonstrate safety in a clinical setting.[27] The FDA has outlined key actions that they intend to take in regards to its regulatory framework, including promoting device transparency to users and developing real-world performance monitoring pilots.[27]

### Deployment

It is important to consider how a model will be incorporated into existing clinical workflows, with disruption kept to a minimum. Ideally this should be considered from the outset to ensure that what is designed is actually useful. Interoperability is a key determinant in ensuring that models may be integrated across different software. Interfacing within electronic health record systems can be challenging, although vendors are moving towards accommodating this.[28] Input from a software engineer is likely to be of value at this stage. Good performance at the time of deployment does not guarantee that the model will continue to perform well, so measures should be put in place for detecting and responding to changes.

It is important to educate and train the workforce who will use the model, stating clearly what the model should be used for and what its limitations are.[29] There may be resistance to the introduction of new technology and, thus, it should be explained as openly and clearly as possible. Clinicians should be involved in the development process and feedback should be sought throughout.

## Conclusion

The development of AI to improve patient care and to enhance clinical research presents great potential and has attracted widespread attention from researchers and the public in recent years. However, this has been accompanied by the emergence of a number of potential barriers to adoption, including ethico-legal concerns. Developing clear guidance for researchers to appropriately frame and answer AI-related research questions remains a clear research priority for the medical field. ■

## Acknowledgements

We thank Prof Parashkev Nachev for his thoughtful comments and suggestions on our manuscript.

## References

1 Melton M. Babylon health gets $2 billion valuation with new funding that will help it expand in US. *Forbes* 2019. www.forbes.com/sites/monicamelton/2019/08/02/babylon-health-gets-2-billion-valuation-with-new-funding-that-will-help-it-expand-in-us [Accessed 01 September 2021].

2 Browne R. AI pharma start-up BenevolentAI now worth $2 billion after $115 million funding boost. *CNBC* 2018. www.cnbc.com/2018/04/19/ai-pharma-start-up-benevolentai-worth-2-billion-after-funding-round.html [Accessed 01 September 2021].

3 Tozzi J. Amazon-JPMorgan-Berkshire Health-Care venture to be called Haven. *Bloomberg* 2019. www.bloomberg.com/news/articles/2019-03-06/amazon-jpmorgan-berkshire-health-care-venture-to-be-called-haven [Accessed 01 September 2021].

4 Academy of Medical Royal Colleges. *Artificial Intelligence in healthcare*. AoMRC, 2019. www.aomrc.org.uk/reports-guidance/artificial-intelligence-in-healthcare [Accessed 02 March 2021].

5 Park Y, Jackson GP, Foreman MA *et al*. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 2020;3:326–31.

6 Vollmer S, Mateen BA, Bohner G *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927.

7 Moons KGM, Altman DG, Reitsma JB *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.

8 Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.

9 Smeaton J, Christie L. *AI and healthcare*. UK Parliament, 2021. https://post.parliament.uk/research-briefings/post-pn-0637 [Accessed 18 April 2021].

10 Garud R, Tuertscher P, de Ven AHV. Perspectives on innovation processes. *Acad Manag Ann* 2013;7:775–819.

11 Kather JN, Krisam J, Charoentong P *et al*. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med* 2019;16:e1002730.

https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002730 [Accessed 01 September 2021].

12 Hollon TC, Pandian B, Adapa AR *et al*. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med* 2020;26:52–8.

13 Hilton CB, Milinovich A, Felix C *et al*. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *Npj Digit Med* 2020;3:1–8.

14 Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open* 2020;3:e1918962.

15 Artzi NS, Shilo S, Hadar E *et al*. Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* 2020;26:71–6.

16 Commandeur F, Slomka PJ, Goeller M *et al*. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. *Cardiovasc Res* 2020;116:2216–25.

17 Arora A. Conceptualising artificial intelligence as a digital healthcare innovation: an introductory review. *Med Devices Auckl NZ* 2020;13:223–30.

18 Fry A, Littlejohns TJ, Sudlow C *et al*. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *Am J Epidemiol* 2017;186:1026–34.

19 Liu X, Faes L, Kale AU *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271–97.

20 Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *ArXiv* 2016:160204938v3 [cs.LG]. http://arxiv.org/abs/1602.04938 [Accessed 01 September 2021].

21 Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak* 2012;12:8.

22 Poplin R, Varadarajan AV, Blumer K *et al*. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;2:158–64.

23 Topol EJ. What's lurking in your electrocardiogram? *Lancet* 2021;397:785.

24 Marcel S, Millán JDR. Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. *IEEE Trans Pattern Anal Mach Intell* 2007;29:743–52.

25 Mindsync. *Is high computing power a roadblock in the path to AI systems deployment? Probably not*. Mindsync, 2021. https://medium.com/mindsync-ai/is-high-computing-power-a-roadblock-in-the-path-to-ai-systems-deployment-probably-not-6c7c5772e7e2 [Accessed 01 September 2021].

26 Hao K. The computing power needed to train AI is now rising seven times faster than ever before. *MIT Technology Review* 2019. www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before [Accessed 01 September 2021].

27 Center for Devices & Radiological Health. *Artificial intelligence and machine learning in Software as a Medical Device*. US Food and Drug Administration, 2021. www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [Accessed 10 July 2021].

28 Davenport TH, Hongsermeier TM, Cord KAM. *Using AI to improve electronic health records*. Harvard Business Review, 2018. https://hbr.org/2018/12/using-ai-to-improve-electronic-health-records [Accessed 01 September 2021].

29 Arora A. Shooting from the hip into our own foot? A perspective on how artificial intelligence may disrupt medical training. *FHJ* 2020;7:e7–8.

**Address for correspondence: Mr Anmol Arora, School of Clinical Medicine, University of Cambridge, Hills Road, Cambridge CB2 0SP, UK.**
**Email: aa957@cam.ac.uk**
**Twitter: @AnmolArora_98**