Thinking Critically about AI in Healthcare

Jessica Morley
Oxford Internet Institute, University of Oxford, 1 St. Giles' OX1 3JS
May 2023

Introduction

The idea that 'Artificial Intelligence' (AI) might be useful for healthcare dates back to at least the 1940s, and enthusiasm for the idea has increased steadily since the 1960s with the sequential development of: Bayesian methods; pattern recognition methods; machine learning methods (in particular deep learning methods); natural language processing; and most recently the introduction of 'foundation models' [1–5]. The list of potential Type A (Medical) and Type B (Administrative)[6] uses of AI in healthcare is now extremely long, covering everything from: risk prediction (i.e., the likelihood of a patient developing a specific condition)[7]; interactive note taking; patient chatbots; in clinic or bedside clinical decision support[2]; drug discovery[8]; cancer classification; image interpretation; disease diagnoses; disease prognoses; treatment identification[9]; and patient-doctor communication[10]. In short, it is proposed, that AI can theoretically help with almost any clinical task. The current level of enthusiasm surrounding AI for healthcare is not, therefore, unjustified[11]. It is clear that it has tremendous potential[11]. Consequently, it is not surprising that policymakers, journalists, and tech-enthusiasts alike are increasingly waxing lyrical about the myriad potential benefits of AI, including: improving the accuracy of diagnosis[12]; improving the reliability of decision-making; improving efficiency of tests[1]; improving compliance with evidence-based care pathways[13]; reducing costs[14]; reducing medication errors[15]; and – of course – making medicine more Precise, Personalised, Preventive, Predictive, and Participatory[16–20] with the twin hopes of achieving the triple aim (reducing per capita costs whilst improving the experience and outcomes of care)[21] and creating a 'learning healthcare system'[22]. The problem is that enthusiasm for all these potential benefits is currently masking the very real limitations, implementation challenges, ethical considerations, patient safety risks [23], and regulatory hurdles that are – in the vast majority of cases – preventing any of the above theorised benefits from being realised in practice[9,24]. The reality of AI for healthcare, stripped of its overly flattering hype, remains woefully underexamined[25]. This then is the purpose of this guide to thinking critically about AI for healthcare, to encourage all interested stakeholders and bystanders to think seriously about the technical, ethical, regulatory, legal, and sociocultural implications of, and barriers to, AI for healthcare[26]. Highlighting a range of issues from the lack of causal and semantic understanding of AI[24]; the challenges associated with validation[2]; the extensive computational and training needs[10]; domain complexity and temporality[4]; privacy concerns; lack of resilience[9]; and (undesirable) transformative effects, and presented as an annotated bibliography with brief summaries, I hope that it is both accessible and useful to all those who stumble across it.

What can AI be used for in healthcare?

Broadly speaking the tasks that AI can be put to in healthcare fall into one of four categories: Prediction, Classification, Association, and Optimisation[27]. Prediction tasks involve using historical data to predict the likelihood of future events, for example, using historical patterns in EHR to identify risk factors for developing specific diseases and using this information to predict the likelihood of one individual developing said disease[7]. Classification tasks involve the recognition of anomalies, for example, recognising the presence or absence of disease in an image, or pathology test. Association tasks (also sometimes referred to as prediction tasks involving the extraction of previously unknown knowledge[28]) are often research tasks such as the identification of new symptoms or risk factors of disease or drug discovery. Finally, Optimisation tasks are typically administrative involving, for example, the scheduling of appointments or the rostering of staff. Most current examples of 'AI products' (including those approved as medical devices in the USA and the EU[29]) are designed to complete relatively narrowly-defined prediction or classification tasks[30], deployed as some form of clinical decision support tool[31], with the aims of improving the accuracy of diagnosis; improving the reliability of decision-making; improving efficiency of tests and therapies; and enabling further research[1]. Of these narrow tasks, by far the most common is image classification, i.e., classifying X-Ray, MRI, CT, or other 'scan' images into normal or abnormal categories. More recently, the development of 'Foundation Models' (including Large Language Models) have prompted theoretical discussions about the development of 'General Artificial Medical Intelligence' with the hope being that the far greater flexibility of Foundation Models, including their ability to adapt to new tasks for which they have not been trained, might enable a move beyond the current one-model-per-task modus operandi[2]. This is, however, currently a purely theoretical hope.

Paper	Why is it useful?
Awaysheh, Abdullah, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan, and Kurt L. Zimmerman. 'Review of Medical Decision Support and Machine-Learning Methods'. Veterinary Pathology 56, no. 4 (July 2019): 512–25.	This paper provides a very detailed overview of the specific algorithms most used in AI models for healthcare. Specifically, it covers the most common machine-learning algorithms: naïve Bayes, decision trees, and artificial neural networks. These are all examples of algorithms trained using 'supervised learning' methods. The paper also covers unsupervised and reinforcement learning methods but does not give examples of the specific algorithms that fall under these umbrella headings.

Baalen, Sophie, Mieke Boon, and Petra Verhoef. 2021. 'From Clinical Decision Support to Clinical Reasoning Support Systems'. *Journal of Evaluation in Clinical Practice* 27(3): 520–28.	This paper explains the purpose of Clinical Decision Support Systems (or Software) that use artificial intelligence in the 'back-end' to mimic the clinical reasoning (or epistemological reasoning) process typically followed by clinicians. Referring to these types of clinical decision support systems as 'algorithmic clinical decision support), the authors state that they can be used to help answer questions such as 'What are the chances that a patient with symptoms x,y,z has disease A? Or disease B?' or 'How likely is it that treatment T will be effective for a patient with symptoms x,y,z?'
De Silva, Daswin, and Damminda Alahakoon. 2022. 'An Artificial Intelligence Life Cycle: From Conception to Production'. *Patterns* 3(6): 100489.	This paper outlines the four primary capabilities of AI: Prediction (e.g., risk prediction or likelihood of a person developing a specific disease), Classification (e.g. presence or absence of disease), Association (e.g., identification of new risk factors or drug discovery), and Optimisation (e.g., administrative tasks such as surgery or appointment scheduling). These capabilities and tasks can be used to identify the specific algorithm that would be most suitable. For example, artificial neural networks (ANN) are most suited to Prediction tasks, Naïve Bayes algorithms to Classification, Gaussian Mixture Models for Association, and Generative Adversarial Networks for Optimisation tasks. The paper includes a helpful taxonomy showing these connections.
Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. 'How to Develop Machine Learning Models for Healthcare'. Nature Materials 18, no. 5 (May 2019): 410–14	This paper provides more detail on the types of prediction tasks that might be completed by AI models. More specifically, it breaks prediction tasks into two categories: learning from humans and enabling extraction of previously unknown insights. Examples of 'learning from humans' include screening (e.g., looking in Electronic Health Records for known signals of risk). Examples of extracting previously unknown knowledge include the identification of new risk factors (e.g., looking for connections between clinical risk factors and social risk factors). Combined these two types of prediction task can increase the accuracy and efficiency of diagnosis, and improve both diagnosis and prognosis (e.g., by identifying a wider range of symptoms and so making it easier to diagnose – particularly rare conditions).
Hall, Peter S., and Andrew Morris. 2017. 'Predictive Analytics and Population Health'. In *Key Advances in Clinical Informatics*, Elsevier, 217–25.	This paper provides a helpful definition of predictive analytics in healthcare stating that it is 'the use of statistics, epidemiology, data mining, machine learning, and artificial intelligence techniques to identify the likelihood of future events based on historical data.' It goes on to explain that the primary aim of predictive analytics is to determine the risk of a patient developing a specific condition, or to determine their likely reaction to a specific treatment. The paper also provides a brief overview of the history of predictive analytic methods, provides some examples of the use of predictive analytics in clinical care, and offers an introduction to new methods of prediction.

Moor, Michael et al. 2023. 'Foundation Models for Generalist Medical Artificial Intelligence'. *Nature* 616(7956): 259–65.	This paper introduces Foundation Models, including Large Language Models, and their potential uses in healthcare – particularly their potential to act as 'General Medical Artificial Intelligence.' It is the potential 'general' capability that is most unique. Most other AI models are only capable of completing relatively narrow, pre-defined tasks. Foundation models could, in theory, adapt to other tasks without training. As a consequence of this 'generalist' capability, the list of potential uses of Foundation Models is long including: interpretation of radiology reports, augmentation during surgical procedures, real time clinical decision support, interactive note taking, patient-facing chatbots, and more.
Muehlematter, Urs J, Paola Daniore, and Kerstin N Vokinger. 2021. 'Approval of Artificial Intelligence and Machine Learning-Based Medical Devices in the USA and Europe (2015–20): A Comparative Analysis'. The Lancet Digital Health 3(3): e195–203.	This paper provides an overview of approved medical devices that make use of AI models in both the USA and Europe, it is a useful paper for understanding what tools are available for use 'on the frontline' as compared to in research.
Will ChatGPT Transform Healthcare?' 2023. *Nature Medicine* 29(3): 505–6.	This paper also discusses the potential use cases of foundation models, specifically Large Language Models (like ChatGPT) but highlights their more qualitative potential uses such as a potential ability to help facilitate conversations between doctors and patients where language or communication skills might act as a barrier.
Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. 'Predicting the Future — Big Data, Machine Learning, and Clinical Medicine'. *New England Journal of Medicine* 375(13): 1216–19.	This paper again classifies the tasks that AI models might be able to complete in a clinical setting, focusing specifically on diagnosis, prognosis, and both radiology and pathology.
Reisman, Y. 1996. 'Computer-Based Clinical Decision Aids. A Review of Methods and Assessment of Systems'. *Medical Informatics* 21(3): 179–97.	This is an old but classic paper highlighting the long history of the idea that artificial intelligence might be used to complete clinical tasks. It includes a long list of reasons for using computer models, including: the improving the accuracy of diagnosis; improving the reliability of decision-making; improving efficiency of tests and therapies; improving the understanding of the structure of medical knowledge; improving the training of diagnostic techniques; and enabling further research.

Why is its use so appealing?

At a higher level of abstraction, beyond the very specific 'tasks' that AI may be used for, the rhetoric used by policymakers and others wishing to encourage and justify the use of AI in healthcare relies heavily on 'systems biology[17],' 'precision medicine[16]' and the idea of a 'learning healthcare system'[22]. The idea is that by using AI to integrate clinical, multi-omic, and epidemiological data and so develop a more detailed understanding of the mechanisms, prognosis, diagnosis, and treatment disease. This information can then be made available to healthcare providers at the point of care via clinical decision support systems (or software) and so make medicine more 'precise' (i.e., more predictive, preventative, personalised, and participatory). Then, finally, by monitoring compliance with clinical decision support recommendations (pathway compliance)[13], as well as outcomes of specific treatments, and demand for specific 'clinics' in certain areas the healthcare system can 'learn' how to deliver care in the most cost effective and efficient way without compromising on outcomes[14]. This, it is hoped, would help to achieve the 'Triple Aim' of improving outcomes and experience of care whilst reducing per capita costs. Traditional statistical models cannot cope with the volume of data involved in such complex analytics, nor the number of variables, and so AI is central to this 'vision.'

Paper	Why is it useful?
Blaser, R. et al. 2007. 'Improving Pathway Compliance and Clinician Performance by Using Information Technology'. *International Journal of Medical Informatics* 76(2–3): 151–56.	This paper is not specifically about AI, but it is a useful introduction to the idea that Clinical Decision Support Software (CDSS) might be used to reduce unwarranted variation in care, by ensuring clinicians have access to the 'right information, about the right patient, at the right time.' Traditionally, this type of CDSS would have involved 'pre-trained' knowledge and act most like a computerised form of a flowchart. Increasingly these traditional passive forms of CDSS are being replaced by active or algorithmic forms of CDSS which have an AI-model running in the background to determine the appropriate advice to provide the clinician, rather than a simple 'if this, then that' set of rules.
Ahmed, Zeeshan. 2020. 'Practicing Precision Medicine with Intelligently Integrative Clinical and Multi-Omics Data Analysis'. *Human Genomics* 14(1): 35.	This paper makes explicit the oft-implied link between AI and precision medicine. Precision Medicine (sometimes referred to as P4 Medicine or, given that it is derived from Systems Biology, Systems Medicine) is the idea that by identifying the exact mechanisms of disease – their spread, their impact on bodies, individual people's susceptibility and responsiveness etc. – down to the most minute level involving thousands of variables, it might be possible to make medicine more Predictive, Preventative, Personalised, and Participatory. By making medicine more 'precise' in this manner, it is hoped that it will be made cheaper, and more effective. Progress in precision medicine has thus far been hindered by the limitations of traditional statistical models for integrating multiple sources of data. AI models, in contrast, are capable of integrating clinical, multi-omic, and epidemiological data making the possibility of precision medicine far more realistic.
Bousquet, Jean et al. 2011. 'Systems Medicine and Integrated Care to Combat Chronic Noncommunicable Diseases'. *Genome Medicine* 3(7): 43.	This paper provides a comprehensive overview of systems biology, the theory that underpins the hopes of Precision or P4 medicine, describing it as 'the use of the power of computational and mathematical modelling to enable understanding of the mechanisms, prognosis, diagnosis, and treatment of disease.'
Deeny, Sarah R, and Adam Steventon. 2015. 'Making Sense of the Shadows: Priorities for Creating a Learning Healthcare System Based on Routinely Collected Data'. *BMJ Quality & Safety* 24(8): 505–15.	This paper provides an overview of another key concept in the rhetoric used by policymakers and others to justify the adoption of AI in healthcare: the Learning Healthcare System. The idea behind the learning healthcare system is that by analysing patterns in routinely collected 'administrative' data, the healthcare system can 'learn' how to deliver care in a way that is more cost efficient and more efficacious. For example, it might 'learn' where clinics for (e.g.,) diabetes are best located, or it might learn that one medicine is more cost effective than another and

	should be made the preferred option in a treatment guideline.
Shapiro, D W, R D Lasker, A B Bindman, and P R Lee. 1993. 'Containing Costs While Improving Quality of Care: The Role of Profiling and Practice Guidelines'. *Annual Review of Public Health* 14(1): 219–41.	This is an old paper that makes clear the link between cost containment and clinical practice guidelines. Clinical Practice Guidelines are the 'original' clinical decision support tool, now being replaced by AI-based systems, and so this paper also makes clear the link between AI and cost containment.

How is it developed?

The development process for AI models intended to be used in clinical settings is significantly more complicated than the standard process used for models that are either intended for use in lower-risk industries, or intended for research-only processes in the clinical domain[26]. This is not to say that the life cycle of AI-for-health models is entirely unique. Many of the stages of model development from problem identification, to ethical review, data storage, data pre-processing, data curation, model training, model evaluation, model augmentation, deployment, and post-deployment surveillance are standard regardless of the intended end-use[27]. Similarly, concerns related to privacy, cybersecurity, trust, Explainability, robustness, usability and overfitting or underfitting[27,28] are broadly applicable. However, there are several additional considerations and stages that *are* unique to the use of AI in healthcare. Starting with healthcare data itself, which is often messy, siloed, and poorly understood by those outside the clinical domain, there are issues related to non-atomicity, temporarily, irregularity, variation in density, multi-modality, missingness, lack of labels, complexity in interpretability, and the need to garner patient and public trust for its use (something which has proven difficult in the past). If insufficient attention is not paid to these additional complexities, and how they will be handled by the model in 'training,' then model performance is likely to suffer [30,32,33]. Further on in the development pipeline, arises the need to conduct clinical trials for the purpose of testing not only model 'accuracy' but also clinical efficacy, and value for money. This is a task that can sometimes be hampered by a lack of guidelines on what is required, cost, and the tendency for AI developers to overly rely on 'common sense' baselines when testing for efficacy (i.e., 'can it beat the human at task X') which are not always readily available in clinical contexts, but is essential if the AI model is to be successfully regulated as a medical device[26,28]. Following this, there are also considerations related to usability, such as whether the model is capable of accurately mimicking all stages of a clinician's decision-making process, how it will fit with the *entire* clinical workflow – not just the specific clinical task it is being used for -, how it will handle the need to be flexible (i.e. to provide differential diagnoses), whether it will be adaptable to different local clinical circumstances, how it will handle important trade-offs such as that between computational efficiency (in often resource-stretched environments) and accuracy or between sensitivity and specificity, and how it will be integrated with existing clinical systems such as EHRs[32,34]. Finally, there is a need to conduct post-market surveillance, as is common in drug development, when the impact of the model is evaluated post-implementation to identify any issues related to patient safety or unintended effects that were not identified and 'designed out' during the development process. This will require healthcare providers to put in place procedures for error reporting and monitoring that may not be typical for IT-systems. All such factors need to be considered in the project management[35] of any AI for health development process and yet are commonly ignored or not reported in the literature.

Paper	**Why is it useful?**
Awaysheh, Abdullah et al. 2019. 'Review of Medical Decision Support and Machine-	This paper provides a brief overview of the history of AI in medical decision making, a deep-dive into the most used models for 'classification' tasks (those most common in decision support) –

Learning Methods'. *Veterinary Pathology* 56(4): 512–25.	including Bayes classifiers, neural networks, and decision trees. It concludes with a discussion of the stages involved in assessing the data quality of the underlying dataset.
De Silva, Daswin, and Damminda Alahakoon. 2022. 'An Artificial Intelligence Life Cycle: From Conception to Production'. *Patterns* 3(6): 100489.	This paper provides a detailed description of the 19-stages of AI production from design to deployment. It is highly practical and comprehensive, based on the experience of the authors experience at the Centre for Data Analytics and Cognition (CDAC), La Trobe University, Bundoora, VIC, Australia. It is particularly useful for thinking through the 'risk assessment' tasks at each stage of the life cycle, and highlights the importance of thinking through all the sociocultural implications of AI related to privacy, cybersecurity, trust, Explainability, robustness and usability. Usability features rarely in other similar papers, and yet it is an extremely important aspect of the efficacy and safety of AI models.
Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. 2019. 'How to Develop Machine Learning Models for Healthcare'. *Nature Materials* 18(5): 410–14.	This paper provides a detailed description of the development pathway for ML models used in healthcare specifically for clinical decision support. It covers the issues that arise at the development, validation, and implementation stages of ML development.
Ngiam, Kee Yuan, and Ing Wei Khor. 2019. 'Big Data and Machine Learning Algorithms for Health-Care Delivery'. *The Lancet Oncology* 20(5): e262–73.	This paper provides a detailed overview of the development process for AI model development, including the importance of evaluating the efficacy of AI models via clinical trials. The authors are particularly keen to point out that the training and evaluation needs of an AI model depend on whether the model is going to be used for research or clinical purposes. Of particular note is the paper's reference to the impact of ML tools on the process of human decision-making, stressing that it is essential to consider the effect of the model's use on the overarching clinical pathway, not just the specific task it is use for. For this reason, the paper places particular emphasis on the need to consider and evaluate human-machine interaction on model performance and patient safety.
Osop, Hamzah, and Tony Sahama. 2019. 'Systems Design Framework for a Practice-Based Evidence Approached Clinical Decision Support Systems'. In *Proceedings of the Australasian Computer Science Week Multiconference*, Sydney NSW Australia: ACM, 1–6.	This paper provides a high-level summary of what clinical decision support systems should be capable of achieving. It states that CDSS should: provide patient-specific recommendations that are relevant to the clinical situation at hand; provide a holistic overview of the patient; be intuitive and easy to use; integrate with EHR systems; offer recommendations that are clearly visible on screen; include explanations for their recommendations; and align with the daily working practices of the specific healthcare provider – including preferences regarding schedule.
Prausnitz, Stephanie et al. 2023. 'The Implementation Checklist: A Pragmatic Instrument for Accelerating RESEARCH-TO-IMPLEMENTATION Cycles'. *Learning Health*	Based on a systematic review, this paper provides a one-page implementation planning checklist incorporating core concepts of existing frameworks related to evidence-based healthcare innovation cycles. It is not as detailed as other papers, and misses some key nuances, but its question-based approach is useful. It covers 'timing and people'; 'data and technology stakeholder engagement and planning' and project management. The less technical approach helps highlight the importance of considering factors such as which specific clinical changes will be brought about by the implementation of a new AI model or tool.
Xiao, Cao, Edward Choi, and Jimeng Sun. 2018. 'Opportunities and Challenges in Developing Deep Learning Models Using	This paper provides a technical discussion of the very specific considerations that arise from the use of EHR data in AI model development, and how potential challenges can be tackled during

Electronic Health Records Data: A Systematic Review'. *Journal of the American Medical Informatics Association* 25(10): 1419–28.	the development process. Specifically, it focuses on considerations related to temporality and irregularity, multi-modality, labelling, and interpretability. Of particular note, is the point made about the challenges involved in identifying true signals from noise in EHR data due to the complex association between clinical events – highlighting the risks of spurious correlations and confounding. In addition, the paper highlights the fact that patient records are not uniform in terms of data density since events are irregularly sampled. The paper concludes with a warning that unless these challenges are tackled, model performance is likely to suffer.
Zikos, Dimitrios. 2017. 'A Framework to Design Successful Clinical Decision Support Systems'. In *Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments*, Island of Rhodes Greece: ACM, 185–88.	Taking a more clinically led to the description of the AI Model development process, this paper first highlights the challenges that are unique to clinical data: non-atomicity (each piece of healthcare data mist be assessed in combination with other data sources); cognitive flexibility (i.e., the importance of presenting differential diagnosis); longitudinally; and shareability (the importance of shared and multidisciplinary decision making). In the second half of the paper, the authors set out a seven-principle framework for the development of CDSS so that it mimics the human clinical decision-making process. It notes that CDSS should provide recommendations with longitudinal insight; should 'know' the time when decisions will be made (i.e., at what point in a clinical pathway will it be used); should provide predictions in a dynamic manner (adaptable to local circumstances); should be outcomes-based; should model a-priory known interactions between clinical attributes; and must be designed with trade-offs (for example between computational efficiency and loss of accuracy in mind).

What are its limitations?

As much as AI enthusiasts are keen to extol its many potential benefits for healthcare – of which there are many – it is important not to get carried away and to remain realistic, balancing hyperbolic claims about AI's achievements with grounding reminders of its limitations. Many limitations are model or use-case specific, but there are some limitations that are generalisable, including the fact that current models are limited in terms of robustness – often not resilient to tiny perturbations[9]- ; that complex models raise privacy and security problems given the volume of data required for training[9]; that the computational requirements can be excessive and difficult to meet[2]; that validation is difficult as is reproducibility[10]; that there are ongoing issues to do with bias that cannot be easily resolved with technical 'solutions'[9]; that ML models cannot reveal insights into causality and can sometimes become overly reliant on spurious correlations[24]; that diseases are always progressing and changing over-time in non-deterministic ways and yet most models assume that (e.g.,) associations between variables stay static[4]; and that there are typically only a few patients with a particular presentation of a disease which can cause issues in terms of accuracy and generalisability[4].

Does it actually work?

When assessing the value, utility, or indeed efficacy of an AI model it is important to remember that building an accurate or high-performing AI model and writing about it in an academic publication, is not the same as building an AI model that is ready for deployment in a clinical system. Moving from 'the lab' to 'the clinic' is a key part of translational medicine and yet very few AI models have yet successfully made the leap across this 'chasm.' One of many reasons for this is the fact that, despite an extremely dense body of literature being available on both the why and the how of evaluating the efficacy, safety, and accuracy of AI models[36–41], the current state of evidence supporting the claims by AI developers is extremely poor. Whilst isolated studies, show that AI models are capable of providing textual answers to patient questions and to medical-exam questions that match or outperform answers provided by human clinicians[42,43], a number of systematic reviews have revealed considerable limitations with the existing literature evaluating the performance of AI models. There is a tendency for the literature to focus solely on 'technical evaluations' (e.g., reporting the ROC curve of an algorithm)[28,44] rather than broader clinical evaluations that might also include an evaluation of the model's clinical efficacy i.e., the impact of model use on patient outcomes (e.g., via an Randomised Controlled Trial)[40,41,45], broader patient needs (such as autonomy)[46], or effective resource use[47]. What's more, even when clinical evaluations are conducted, these are often still done only in 'the lab' rather than in 'the real world' and are reported poorly[48], with reported evaluations omitting key details such as what data the model was trained on[49] – and so severely limiting opportunities for replication[50]; relying on spin practices and inflating results (either purposefully or unknowingly)[51,52]; and being at high risk of bias[53]. In short, before any claims about the efficacy or utility of AI models and their ability to enhance the capabilities of human healthcare practitioners can be taken seriously, far more attention needs to be paid to the generation of high-quality evidence[54]. Otherwise, rushed implementation and deployment might lead to increased strain on the healthcare system, undue stress to patients, and possible harm from mis or missed diagnosis (or other error)[40].

Paper (Why and How of Evidence Generation)	**Why is it useful?**
Calvert, Melanie, Rob Thwaites, Derek Kyte, and Nancy Devlin. 2015. 'Putting Patient-Reported Outcomes on the "Big Data Road Map"'. *Journal of the Royal Society of Medicine* 108(8): 299–303.	This paper makes the key point that if evaluations of effectiveness of any clinical intervention (including data-driven interventions like those based on AI) ignore 'qualitative' indicators like Patient Reported Outcomes, then it is likely that the evaluation will underestimate the impact of the disease on the patients in question and overestimate the effectiveness of the intervention. The authors advocate for the adoption of a more standardised approach to the recording of patient reported outcomes so that these might be more readily included in evaluations.
Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. 'How to Develop Machine Learning Models for Healthcare'. Nature Materials 18, no. 5 (May 2019): 410–14.	This paper concludes its overview of the stages involved in the development of ML models for healthcare with a discussion about the importance of evaluating the model for potential clinical impact – noting that model performance alone is insufficient for creating clinical impact. The authors stress that ML developers must also consider user trust (both under and over-reliance), integration with workflow, and usability.
Doyal, L. 1992. 'Need for Moral Audit in Evaluating Quality in Health Care.' *Quality and Safety in Health Care* 1(3): 178–83.	Like the Calvert paper, this paper also draws attention to the limitations of taking a purely quantitative approach to service and medical intervention evaluation. Doyal points out that 'preventing harm' (and so identifying risks to harm) must also include consideration of a patient's 'human needs' which include needs related to autonomy and the ability to actively participate in society. Although this paper is responding to the introduction of 'audit' to the NHS in the 1990s, it is still highly relevant to evaluations of

	AI-based interventions as many policymakers justify their advocacy for the greater use of AI on the basis of the need to ‘monitor performance’ and eradicate unwarranted variation in care – the aims of audit and feedback. Thus, by extrapolating it becomes clear that the paper makes the point that if the evaluation of AI-based tools focuses exclusively on the ability of tools to ‘police’ clinicians’ ability to stick unwaveringly to practice guidelines, regardless of the patient’s individual circumstances, this would potentially be harmful.
de Hond, Anne A. H. et al. 2022. ‘Guidelines and Quality Criteria for Artificial Intelligence-Based Prediction Models in Healthcare: A Scoping Review’. *npj Digital Medicine* 5(1): 2.	This paper reports an attempt to identify existing guidelines and quality criteria regarding six phases of the AI-based prediction model development, evaluation, and implementation cycle to produce a structured quality assessment framework that can be used for the entire AI lifecycle. It covers the following six stages: (1) preparation, collection, and checking of data; (2) development of the model; (3) Validation of the model; (4) development of the software application that will house the model; (5) impact assessment of the model when contained within the software application; (6) implementation and use of the model in daily healthcare practice. What is produced is a comprehensive list of all the quantitative steps involved in developing, deploying, and using an AI-model intended for clinical use, covering everything from ensuring compliance with data protection law, to the importance of justifying specific model selection, to the importance of external validation, and finally to the need to monitor and audit performance post implementation. These tasks overlap significantly with other similar lists, including those produced earlier (e.g., by Miller) and still lack a discussion of exactly how the tasks should be completed.
England, Joseph R., and Phillip M. Cheng. 2019. ‘Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers’. *American Journal of Roentgenology* 212(3): 513–19.	This paper provides an excellent overview of what readers should look out for when analysing papers claiming to have validated an image-recognition algorithm. It takes pains to highlight the fact that ‘accuracy’ is not a sufficient guarantee that an algorithm will work in clinical practice, and that different statistical measures of ‘accuracy’ can be misleading. It points out that whilst the ROC curve is the most commonly used performance metric – this doesn’t provide insight into the region of the curve where sensitivity and specificity are balanced the area that most clinicians may wish to examine) and so a contingency table containing: true-positive, true-negative, false-positive, false-negative rates and derivatives of these measures, such as sensitivity, specificity, PPV, NPV, and likelihood ratio should also be included, alongside a summary statistic such as the ROC AUC. It also provides insight into the evaluations statistics that might be used for non-classification ML models (e.g., prediction models –

	such as mean, absolute error, mean squared error, and root-mean-square error. The paper concludes that authors should provide as many metrics as necessary to describe the strengths and weaknesses of a given algorithm and that – where possible – these evaluation metrics should be reported with confidence intervals or a measurement of statistical significance (particularly when making comparisons between algorithms or between an algorithm and a human).
Lisboa, P.J.G. 2002. 'A Review of Evidence of Health Benefit from Artificial Neural Networks in Medical Intervention'. *Neural Networks* 15(1): 11–39.	Although this paper might appear out of date, it provides a very clear description of the steps involved in evaluating complex models intended for use in healthcare. It covers all steps from clarifying the purpose of the study; to validating the performance of the model; to benchmarking the performance against a suitable alternative; to testing the robustness of the performance; and finally, to conducting comparative trials. It concludes that following this process must become 'the norm' as currently many of the claims of 'prototype studies' are not sufficiently robust – a fact that, sadly, remains true more than twenty years later.
Liu, Vincent X, David W Bates, Jenna Wiens, and Nigam H Shah. 2019. 'The Number Needed to Benefit: Estimating the Value of Predictive Analytics in Healthcare'. *Journal of the American Medical Informatics Association* 26(12): 1655–59.	This paper takes a different approach to evaluating AI models – specifically complex predictive models. Instead of focusing on 'technical performance,' it instead focuses on health economics, and highlights both the need to and the difficult of balancing the resources used and benefits gained by developing and deploying a particular predictive model. The authors make the argument that learning how to assess 'value'(i.e., assess this balance) is critical to the adoption of safe, effective, and sustainable predictive models. The authors go on to elucidate 2 components of a 'value framework' for evaluating predictive models: (1) the number needed to screen (the number of patients the model must flag to identify 1 true positive); and (2) the number needed to treat (estimates the number of true positive patients that must be treated for 1 patient to benefit). They argue that the product of NNS and NNT produces 'the number needed to benefit' and that this number contextualised with the costs of screening and treatment can help highlight the costs and benefits of actions that result from responding to a model's predictions. On the surface, this may seem a relatively simple proposal, but it is highly useful for highlighting the fact that a more comprehensive approach to evaluating a model's utility is needed beyond pure 'accuracy' metrics.
Mahadevaiah, Geetha et al. 2020. 'Artificial Intelligence-based Clinical Decision Support in Modern Medical Physics: Selection, Acceptance, Commissioning, and Quality Assurance'. *Medical Physics* 47(5).	This paper, unusual in the literature, focuses on the steps involved in evaluating an AI-based tool once it has been developed and when it is to be implemented into a clinical setting. It, therefore, covers different stages of the lifecycle, focusing on how clinical providers should select between different potential models or solutions, how acceptance testing should be

	conducted, how models should be commissioned and implemented, and how quality assurance should be conducted post implementation. The Quality Assurance section of the paper is the most detailed and the most novel, noting that it is crucial for healthcare providers to have systems in place for monitoring and evaluating both efficiency and efficacy, as well as malfunctions, and both external (context) and internal (model) drift that might impact model performance over time. It is, for this reason, one of only a few papers that point out the need for local validation studies to be conducted in addition to large general validation studies.
Miller, Perry L. 1986. 'The Evaluation of Artificial Intelligence Systems in Medicine'. *Computer Methods and Programs in Biomedicine* 22(1): 3–11.	An old but classic and still extremely relevant paper that discusses the three general stages at which evaluation of an AI-model should be conducted: (1) the subjective assessment of the research contribution of a developmental system; (2) the validation of a system's knowledge prior to possible clinical use; and (3) the evaluation of the clinical efficacy of an operational consultation system. Going into more detail on validation, Miller elucidates that this must involve static validation (examining the system's knowledge when in the lab) and dynamic validation (examining the system's knowledge when in use. Finally, Miller notes that a second task in assuring a system's safety involves exposing it to experimental use in a clinical environment as 'even if the system's knowledge is indeed accurate, complete, and consistent, it will be of little help if it's clinical interface is faulty.' Although the appearance of more complex models has made the process of validation more complicated, this paper is still an excellent primer on the different stages involved and their importance.
Neves, Mariana R., and D. William R. Marsh. 2019. 'Modelling the Impact of AI for Clinical Decision Support'. In *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, eds. David Riaño, Szymon Wilk, and Annette ten Teije. Cham: Springer International Publishing, 292–97.	This paper too makes an effort to point out that 'prediction accuracy' (or model performance more generally) does not necessarily ensure clinical efficacy or even utility. Assessing whether or not a model will genuinely have an impact on clinical care, must also involve an evaluation of the way in which the prediction (or the model) is intended to interact with other stages of the clinical decision-making process and the proposed benefits (costs, workload, or better decision-making). These impacts, the authors argue, need to be considered before evaluation begins so that a means of testing them can be built into any experimental study.
Nsoesie, Elaine O. 2018. 'Evaluating Artificial Intelligence Applications in Clinical Settings'. *JAMA Network Open* 1(5): e182658.	This paper both makes the case for the importance of evaluating AI-based tools in clinical settings, not just in research settings, and laments the fact that there is currently (and there still is five years later) a dearth of clinical evaluations of AI-based tools. Reasons for the essentiality of clinical evaluation covered by the paper centre on the fact that some performance deficiencies

	may only become apparent once the model is used in a clinical setting because the training datasets will have been careful curated to remove imperfect data samples whereas in the 'real world,' imperfections are more than likely to be present. For example, tbe paper notes that a system trained only on high-quality images might provide incorrect diagnosis when classifying low-quality images or images affected by sheen or other defects present in real-world clinical settings. The authors conclude that without clinical evaluation studies, the implementation of AI might be premature leading to increased strain on the healthcare system, undue stress to patients, and possible death owing to mis or missed diagnosis.
Park, Seong Ho, and Kyunghwa Han. 2018. 'Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction'. *Radiology* 286(3): 800–809.	This paper provides a very pragmatic, and yet detailed, overview of the different methodological approaches available to individuals wishing to evaluate the clinical performance of an AI-based tool. It covers RCTs, noting that these are the 'gold standard of evidence,' as well as observational studies, and prospective before-after studies. The paper includes a highly useful diagram showing clinical trial designs that could be used to evaluate the effect of an AI tool on a patient outcome, it covers both a traditional RCT and a cluster randomised trial (where randomisation is done at the time-period level for pragmatic purposes).

Paper (Quality of Evidence)	**Why is it useful?**
Andaur Navarro, Constanza L. et al. 2023. 'Systematic Review Finds "Spin" Practices and Poor Reporting Standards in Studies on Machine Learning-Based Prediction Models'. *Journal of Clinical Epidemiology*: S0895435623000756.	This paper highlights the tendency for papers reporting on the development of ML-based prediction models to use inappropriate methods and to be incompletely reported. In particular, it focuses on the tendency for papers to rely on 'spin' i.e., language that exaggerates the benefits of ML-based prediction models whilst downplaying the costs, risks, and limitations. 152 studies are included in the systematic review, which found that most articles contained two examples of spin in the Results section, four in the discussion section and at least one in another section. In addition, the Review found that 86/152 studies recommended that their model be used in clinical practice, despite the fact that just 8 of these same models had been externally validated.
Ayers, John W. et al. 2023. 'Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum'. *JAMA Internal Medicine*.	This paper reports the findings of a study where 195 questions posted to Reddit's r/AskDocs forum were put to both a human clinician and a chatbot. The responses were anonymised and analysed by a team of licensed healthcare professionals. Evaluators chose which response was better and judged both the 'quality of information provided' and the empathy or bedside manner provided. Evaluators preferred the chatbots to the clinician responses 78% of the time.

Coiera, E, and HL Tong. 2021. ‘Replication Studies in the Clinical Decision Support Literature-Frequency, Fidelity, and Impact’. *JOURNAL OF THE AMERICAN MEDICALINFORMATICS ASSOCIATION* 28(9): 1815–25.	This paper presents the results of a review designed to assess the frequency, fidelity, and impact of replication studies in the clinical decision support system (CDSS) literature (remembering here that most modern CDSS claims to make use of some form of AI). The review identified 4063 papers matching the search criteria for CDSS research, only 0.3% (or 12 papers) of which were found to be replications. Of these 12 papers, 6 could not reproduce the results, 2 tested variants of the original CDSS and 4 validated measurement instruments. Ultimately, the replication rate was found to be 3 in a thousand studies – an incredibly poor rate, that would be unacceptable in most disciplines and is certainly unacceptable in medicine.
Ge, Wenbo, Christian Lueck, Hanna Suominen, and Deborah Apthorp. 2023. ‘Has Machine Learning Over-Promised in Healthcare?’ *Artificial Intelligence in Medicine* 139: 102524.	This paper describes the many ways in which the performance of AI models is inflated in the current literature, giving the impression that they will perform well in clinical practice when in reality this is unlikely. The inflationary effects covered include: the digital fingerprinting phenomenon (the issue of different samples of the same patient sometimes ending up in different partitions of the data e.g., training and validation datasets); the accuracy paradox (unbiased class distributions resulting in a model that learns the skew of the class distribution and classifying everything as the majority class); and second order effects of the accuracy paradox (when an imbalance of factors such as age, sex, weight etc. in the training dataset is exploited in a similar way to the accuracy paradox).Finally, the paper ends with a discussion of the problem of generalisation, noting that evaluation techniques often have the underlying assumption that the observed data is representative of the population – when, in reality, this assumption rarely holds true.
Gilson, Aidan et al. 2023. ‘How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment’. *JMIR Medical Education* 9: e45312.	This paper reports the results of a study evaluating the performance of ChatCPT on questions within the scope of the US Medical Licensing Exam and to analyse its responses for user interpretability. The model beat the 60% threshold expected of a third-year medical student. However, it was not without fault, with the authors reporting that it made a number of logical, informational, and statistical errors.
Kwan, Janice L et al. 2020. ‘Computerised Clinical Decision Support Systems and Absolute Improvements in Care: Meta-Analysis of Controlled Clinical Trials’. *BMJ*: m3216.	This paper reports the results of a meta-analysis of 122 trials of CDSS embedded in EHR systems. The authors report no significant improvement in clinical outcomes in the subset of 30 trials that included them, and highlighted the fact that there is little to no consistency between trials. The linked editorial (Sarkar, U., & Samal, L. (2020). How effective are clinical decision support systems? The BMJ, 370.) notes that in addition to this overall limited clinical performance, trials of CDSS tend not to report on important context specific implementation metrics such as the number of dismissed alerts, the time required to address recommendations, and clinician satisfaction – further limiting confidence in the quality of the evidence of efficacy and usability.

Plana, Deborah et al. 2022. ‘Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review’. *JAMA Network Open* 5(9): e2233946.	This paper presents the findings of a systematic review of Randomised Controlled trials involving machine learning and finds that relative to the number of ML-based products ‘on the market,’ the number of RCTs involving the same products is startlingly few. More than this, the RCTs that have been conducted were found to be of poor quality, varying considerable in terms of adherence to reporting standards, and with high risks of bias evident in most RCTs reviewed.
Vasey, Baptiste et al. 2021. ‘Association of Clinician Diagnostic Performance With Machine Learning-Based Decision Support Systems: A Systematic Review’. *JAMA network open* 4(3): e211276.	This paper reports the results of a systematic review involving 37 studies evaluating the association between clinician diagnostic performance and the use of ML-based Clinical Decision Support Systems. The review finds no robust evidence to suggest that the use of ML-based clinical algorithms might result in improved clinician diagnostic performance. The authors rightly conclude that the findings indicate a need for caution when it comes to claims regarding the capability of ML-algorithms to positively affect patient care, and further emphasise the need for more high-quality evidence evaluating both the efficacy of ML-based Clinical tools and the factors that affect human-computer interaction.
Yusuf, Mohamed et al. 2020. ‘Reporting Quality of Studies Using Machine Learning Models for Medical Diagnosis: A Systematic Review’. *BMJ Open* 10(3): e034568.	This paper reports the findings of another systematic review assessing the reporting quality of studies developing or validating ML models for clinical diagnosis, focusing especially on what each study reported (or did not report) regarding the demographic and clinical characteristics of the population upon which the model was trained. Amongst other things, the Review found that in more than half of the studies included (54%), it was unclear whether the population included in the study matched, or at least aligned to, the population of the area where the model was intended to be deployed. Ultimately, as the authors state, the review found that studies developing or validating ML-based systems for clinical diagnosis failed to use reporting guidelines and lacked adequate detail for assessment, interpretation, and reproducibility.

How easy is it to implement?

Limitations regarding AI’s robustness and a lack of clinical evidence are not the only ‘inconvenient truths’[55] currently preventing AI from crossing the chasm between ‘the research lab’ and ‘the clinical frontline.’ There are also a number of significant challenges related to its implementation ranging from problems with data quality and access, complexity of clinical workflows, and legacy IT systems and integration challenges, to workforce issues. Despite a number of recent attempts to outline how some of these issues may be overcome to ensure successful AI implementation[25,56], there – as of yet -remains a distinct lack of consensus regarding how exactly this might be achieved. Briefly some of the main issues that need to be overcome relate to data, system integration, complex workflows, and workforce and skill-mix inconsistencies:

- Health data is not always well or consistently structured, does not contain regularly occurring events, and is sometimes plagued with missing values[26,57]. Consequently it must be carefully curated before it can be used for the development of AI (otherwise issues with quality will negatively impact the performance of any model it is used to train[58]) which is a time-consuming and resource-intensive highly-skilled task. In addition, health data is often held in non-integrated siloes, in different formats, with different and inconsistent access rules. In the absence of standards for making these siloes 'talk to each other' this makes gaining access to the volumes of data necessary to train AI models difficult[59]. This is particularly true as the maintenance of patient privacy is essential and effective privacy-preserving measures (such as the use of Trusted Research Environments) are not always suitable or effective in the case of AI development[8,60].
- AI models must be integrated into existing clinical and 'IT' systems if they are to be used at the point of care. Yet, often, these systems are insufficiently robust, reliable, or flexible to support the operation of complex models. In other words, successful implementation of AI relies on a high level of existing technological readiness, and a large number of adequately functioning interconnected subsystems and components[61]. As many healthcare institutions run on 'legacy' IT infrastructure (particularly in publicly funded healthcare systems like the NHS), this 'readiness' is not observable.
- Accurate AI models implemented into robust clinical systems may still fail to be useful if they unnecessarily disrupt existing workflows[62] (and so encourage the development of potentially dangerous workarounds[63]), add additional burden to already complex and overly stretched care pathways, or do not adequately replicate the steps involved in human clinical decision making (i.e., the cognitive workflow of clinicians)[64,65]. These more nuanced considerations rarely feature in the academic literature regarding the development of AI models for healthcare.
- The development of accurate and useful AI models requires a unique mix of skills and experiences: software developers need an understanding of the healthcare system, of medicine, and biology, whilst clinicians need an understanding of how models are developed[24]. Without this knowledge, clinicians may be more vulnerable to automation bias[66] and to misinterpreting results in a way that could result in harm to patients, whilst AI developers may make models fraught with epistemic errors and constraints[67]. In addition, if models are to be capable of reflecting the true complexity of medicine, then they need to be developed by a diverse workforce capable of understanding multiple different 'lived experiences' of health and healthcare and how these differ depending on demographic and socioeconomic factors as well as their interactions . Currently, there is a shortage of individuals who would meet all these criteria[70,71].

Paper – Data	Why is it useful?
Bainbridge, Michael. 2019. 'Big Data Challenges for Clinical and Precision Medicine'. In *Big Data, Big Challenges: A Healthcare Perspective*, Lecture Notes in Bioengineering, eds. Mowafa Househ, Andre W. Kushniruk, and Elizabeth M. Borycki. Cham: Springer International Publishing, 17–31.	This is a fairly UK/NHS-centric paper but it makes the key (and generalisable) point that despite a prevailing believe that 'if only the might of Big Data and Deep Learning were applied to Health and Medicine then the benefits would flow in abundance' – in many cases, the prerequisite existence of ''structured and coded clinical data' is still missing.
Baxter, Sally L., and Aaron Y. Lee. 2021. 'Gaps in Standards for Integrating Artificial Intelligence Technologies into Ophthalmic Practice'. *Current Opinion in Ophthalmology* 32(5): 431–38.	Although the title of this paper implies is rather narrowly focused on the use of AI in Ophthalmology may of the points raised are more generally applicable. In particular, the central argument is that for widespread

	deployment of AI to be possible, there needs to be standards developed to enable seamless and automated transfer of information between different systems including EHRs, imaging systems, and clinical decision support systems, and yet these standards do not currently exist.
Dhindsa, Kiret, Mohit Bhandari, and Ranil R Sonnadara. 2018. 'What's Holding up the Big Data Revolution in Healthcare?' *BMJ*: k5357.	Going into more detail than the Bainbridge paper, the authors explain why current practices around collection, curation, and sharing of health data are currently acting as a major barrier to the development and evaluation of AI for healthcare. The paper concludes that unless better data management practices are adopted and more standardised means of guaranteeing data quality are developed, then AI tools are likely to remain unable to transition successfully form the lab to clinical practice.
Heckman, George A., John P. Hirdes, and Robert S. McKelvie. 2020. 'The Role of Physicians in the Era of Big Data'. *Canadian Journal of Cardiology* 36(1): 19–21.	This paper centres on the argument that clinical decision-making is more than just the application of facts to a case. It is also about contextualisation and the application of 'clinical reasoning' which might not always be reducible to pure numbers or clinical rules. Thus, the authors make the case for seeing AI as an augmentation tool, but not a 'replacement' tool for clinical reasoning.
Kerasidou, Charalampia (Xaroula), Maeve Malone, Angela Daly, and Francesco Tava. 2023. 'Machine Learning Models, Trusted Research Environments and UK Health Data: Ensuring a Safe and Beneficial Future for AI Development in Healthcare'. *Journal of Medical Ethics*: jme-2022-108696.	This is a general comment paper rather than a technical paper (which is really what is needed) but it serves to highlight the point that whilst "Trusted Research Environments" are a useful 'solution' to the challenges presented by the need to simultaneously provide broad data access and protect patient privacy, there are a number of reasons why (current TREs) are unable to support the development of AI models (particularly those based on Machine Learning).
Leyens, Lada, Matthias Reumann, Nuria Malats, and Angela Brand. 2017. 'Use of Big Data for Drug Development and for Public and Personal Health and Care: Leyens et Al.' *Genetic Epidemiology* 41(1): 51–60.	This paper has, at its heart, a clear and simple argument: "more data does not necessarily mean more action." In explaining why this is the case it covers many of the same topics as Bainbridge and Dhindsa, explaining that issues related to data quality can result in significant errors (such as false positives) in any models developed on the inaccurate data. However, its main focus is on the issues that arise from the fact that many current health databases are siloed, unstandardised, unstructured, and unavailable – a problem that is exacerbated by the fact that currently there are no international codes of practice in data management, data access, data querying, or data sharing.
Ngiam, Kee Yuan, and Ing Wei Khor. 'Big Data and Machine Learning Algorithms for Health-Care Delivery'. The Lancet Oncology 20, no. 5 (May 2019): e262–73.	More specific than the other papers, this paper focuses on the very specific data curation challenges faced by developers of ML-algorithms. It explains that data curation is a process of reclassifying data into clinically or logically relevant subgroups that might improve the predictive accuracy of and ML-based algorithm, and that this requires substantial clinical understanding of the data, the nature of the problem, and the performance of the ML methods used. Of particular focus in this

	discussion are the problems of data missingness and ML's preference for 'regularly patterned data' that means often sporadic and seemingly 'random' health datasets are not necessarily 'model development ready.'

Paper – Technology Readiness	**Why is it useful?**
Lavin, Alexander et al. 2022. 'Technology Readiness Levels for Machine Learning Systems'. *Nature Communications* 13(1): 6039.	This paper is fairly long and technical, but it outlines a useful framework defining a principled process for ensuring robust, reliable, and responsible systems that are capable of supporting AI algorithms. It is a generalisable framework, but includes an example of a diagnostic algorithm and how the framework might be applied to its development and implementation.

Paper – Workflow Implications	**Why is it useful?**
Bucur, Anca et al. 2016. 'Workflow-Driven Clinical Decision Support for Personalized Oncology'. *BMC Medical Informatics and Decision Making* 16(S2): 87.	This paper is a little overly-specific and does focus on clinical decision support that may or may not be supported using AI. However, it does make the extremely important point that any decision support tool (including, for example an AI-based predictive model) must be deployed in a way that does not interrupt the clinical workflow.
Goff, Mhorag et al. 2021. 'Ambiguous Workarounds in Policy Piloting in the NHS: Tensions, Trade‐offs and Legacies of Organisational Change Projects'. *New Technology, Work and Employment* 36(1): 17–43.	Again, this paper does not exclusively focus on AI, but the general tendency for 'policies' to be piloted in healthcare systems (in this case the NHS) before they become fully 'adopted.' It makes the crucial, and highly relevant, point that when policies (including technology policies) are implemented in a top-down fashion with little consideration given to the clinical environment or the workflow of the relevant users, it is common for staff to develop 'workarounds' which can both undermine the effectiveness of the policy/technology and, in some instances, introduce patient safety risks.
Rezaei-Yazdi, Ali, and Christopher D. Buckingham. 2018. 'Capturing Human Intelligence for Modelling Cognitive-Based Clinical Decision Support Agents'. In *Artificial Life and Intelligent Agents*, Communications in Computer and Information Science, eds. Peter R. Lewis, Christopher J. Headleand, Steve Battle, and Panagiotis D. Ritsos. Cham: Springer International Publishing, 105–16.	This paper focuses on the utility/trustworthiness of AI models. It argues that, for models to be useful, they need to support the cognitive workflow of the clinicians who will be using them. This means that the way in which they reach an output should – ideally – reflect the mental model of decision-making used by clinicians and the whole process should be interpretable and transparent.

Paper – Workforce	**Why is it useful?**
Buchlak, Quinlan D. et al. 2020. 'Ethical Thinking Machines in Surgery and the Requirement for Clinical Leadership'. *The American Journal of Surgery* 220(5): 1372–74.	This paper summarises the need for clinical involvement in the development of AI systems intended for clinical use. The authors argue, that only by involving clinicians in the development process, can AI developers ensure their models do not waste resources and produce salient outcomes.
Cosgriff, Christopher V, Leo Anthony Celi, and David J Stone. 2019. 'Critical Care, Critical Data'. *Biomedical Engineering and Computational Biology* 10: 117959721985656.	This paper highlights the fact that the methods by which data are explored, processed, harmonised, transformed, and modelled are not currently taught as part of the standard medical education. This can lead to misunderstandings about what is and is not possible to

	achieve with the use of AI models in healthcare, and result in issues related to interpretation of results. Equally, developing accurate and useful AI models requires detailed clinical insight about what questions are relevant to medical care and how the data themselves were generated in practice and this knowledge is not taught to software developers. There is a growing need for cross-over skills and yet there are limited opportunities for individuals to develop this unique mix.
Dullabh, Prashila et al. 2022. ‘The Technology Landscape of Patient-Centered Clinical Decision Support – Where Are We and What Is Needed?’ In *Studies in Health Technology and Informatics*, eds. Paula Otero, Philip Scott, Susan Z. Martin, and Elaine Huesing. IOS Press.	This paper focuses on the ‘softer’ aspects of building useful AI-based models, highlighting that if they are to be used to support the delivery of evidence-based and patient-centred care, that meets the needs of both patients and providers, then it is necessary for AI developers to gain a deeper understanding of patient and provider lifestyles and workflows. Again, this is knowledge that the AI workforce is unlikely to ‘naturally’ possess.
Fridsma, Douglas B. 2018. ‘Health Informatics: A Required Skill for 21st Century Clinicians’. *BMJ* 362: k3043.	This paper presents a less UK-centric argument that informatics should be a fundamental and required skill for clinicians entering the workforce today.
Fosch-Villaronga, Eduard et al. 2022. ‘Accounting for Diversity in AI for Medicine’. *Computer Law & Security Review* 47: 105735.	This paper does not tackle workforce considerations explicitly. However, it makes the point that developing accurate AI models requires a deep understanding of inter-individual differences in how diseases develop and present, including differences related to sex, gender, race, environment, and a wide variety of other socioeconomic factors, and the ways in which all these factors interact. Many of the ‘explanations’ or ‘understandings’ of these differences are unobservable or unknown to those who do not have lived experience of being a person of that (e.g.,) race, sex, or gender. Thus, it is important that AI models are developed by diverse teams who can collectively produce models capable of handling these complexities.
Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. 2014. ‘Automation Bias: Empirical Results Assessing Influencing Factors’. *International Journal of Medical Informatics* 83(5): 368–75.	This paper discusses the concept, causes, and implications of automation bias (i.e., the tendency for clinicians to simply trust a computer because it is assumed that machines are less fallible than humans and thus less likely to make mistakes). It highlights the fact that lower confidence in one’s own ability (including technical ability) increases the likelihood of automation bias.
Scobie, Sarah, and Sophie Castle-Clarke. 2020. ‘Implementing Learning Health Systems in the UK NHS: Policy Actions to Improve Collaboration and Transparency and Support Innovation and Better Use of Analytics’. *Learning Health Systems* 4(1).	This paper highlights the shortage of technical/analytical skills in the healthcare workforce. It focuses on the NHS, but the arguments are generalisable.

Will healthcare practitioners adopt it?

Evidence regarding the 'demand' for clinical AI from healthcare practitioners themselves is mixed. Whilst some studies have found that healthcare practitioners echo the arguments used by policymakers and AI developers and extol the benefits of AI for reducing workload, enhancing efficiency, reducing error, and containing costs[72,73], others have found that healthcare practitioners are reluctant to adopt AI[74] and are concerned about the possible threats posed by AI in terms of clinical skills, capacity, loss of control[75], and professional identity[76]. Furthermore, even when there is evidence supporting healthcare practitioners' willingness to adopt AI, this willingness is conditional. Healthcare practitioners are willing to adopt AI when the model's intended is of clear clinical value[77]; when accompanied by a user-friendly and clinician-centric interface[78,79]; when the model's decision-making process is sufficiently transparent/explainable[80]; when the model has been independently validated and both its potential and limitations have been made clear[81]; when the model allows for contextualisation and clinical judgement[82]; when it clearly improves patient outcomes[83]; and when adequate training in the use of the model has been provided[84]. When these, and several other conditions, are not met then willingness to adopt drops sharply.

Paper	Why is it useful?
Abouzahra, Mohamed, and Dale Guenter. 2022. 'Exploring Physicians' Continuous Use of Clinical Decision Support Systems'. *European Journal of Information Systems*: 1–22	This paper claims that clinicians are more likely to use AI-based clinical decision support systems, if they do not pose a threat to professional identity and can demonstrate an ability to genuinely improve care and so directly benefit patients.
Braun, Matthias, Patrik Hummel, Susanne Beck, and Peter Dabrock. 2021. 'Primer on an Ethics of AI-Based Decision Support Systems in the Clinic'. *Journal of Medical Ethics* 47(12): e3–e3.	This paper sets out the 'conditions of trustworthiness' from the perspective of clinicians, listing: user-friendliness, adequate risk-benefit analysis, appropriate data protection, and evaluation of the system outputs.
Catchpole, Ken, and Myrtede Alfred. 2018. 'Industrial Conceptualization of Health Care Versus the Naturalistic Decision-Making Paradigm: Work as Imagined Versus Work as Done'. *Journal of Cognitive Engineering and Decision Making* 12(3): 222–26.	In this paper, the authors argue that it is important that developers of AI models do not just assume that variability in clinical reasoning is always undesirable and ensure that there is sufficient flexibility in the model to account for clinical judgement (and so variation in treatment) when it is deemed necessary.
Choudhury, Avishek. 2022. 'Factors Influencing Clinicians' Willingness to Use an AI-Based Clinical Decision Support System'. *Frontiers in Digital Health* 4: 920662.	This study found that willingness to use AI depends largely on whether clinicians perceive the model in question to be 'risky.' Going into more detail about the factors that influence risk perception, the authors find that AI developers need to ensure ease of use, be transparent about the model's potential, and pay attention to the design of the model and its enveloping software.
Crigger, Elliott et al. 2022. 'Trustworthy Augmented Intelligence in Health Care'. *Journal of Medical Systems* 46(2): 12.	This paper defines trustworthy as 'dependable and worthy of confidence.' It then provides a framework that the authors intend to be used by clinicians as a mental checklist when deciding whether or not to trust a proposed AI system. The Framework is based around three key questions: does it work? Does it work for patients? Does it improve outcomes?
Jones, Caroline, James Thornton, and Jeremy C. Wyatt. 2021. 'Enhancing Trust in Clinical Decision Support Systems: A Framework for Developers'. *BMJ Health & Care Informatics* 28(1): e100247.	This is a detailed study about the factors that influence clinician trust (an often somewhat amorphous concept) in AI-based clinical decision support systems. It concludes by making three summary recommendations to AI developers: (1) that they should be transparent

	about model content and performance; (2) that they should – as far as possible – avoid the use of black box models; and (3) that they should demonstrate compliance with all relevant and regulatory (codes and standards) frameworks.
Kealey, Edith, Emily Leckman-Westin, and Molly T. Finnerty. 2013. 'Impact of Four Training Conditions on Physician Use of a Web-Based Clinical Decision Support System'. *Artificial Intelligence in Medicine* 59(1): 39–44.	This paper reports the findings of a study which involved the testing of four different models of clinician training in the use of a clinical decision support system. It finds that all training is beneficial, but the most useful is hands-on training rather than lecture based or purely informational training (i.e. the provision of a user-guide).
Nitiéma, Pascal. 2023. 'Artificial Intelligence in Medicine: Text Mining of Health Care Workers' Opinions'. *Journal of Medical Internet Research* 25: e41138.	This study of healthcare practitioner sentiment regarding AI, found that AI tools designed to be used for screening, diagnostic, or treatment purposes were perceived negatively by the clinical community, mostly because there are persistent concerns regarding the accuracy and reliability of AI models as well as concerns regarding their impact on patient privacy.
Rundo, Leonardo et al. 2020. 'Recent Advances of HCI in Decision-Making Tasks for Optimized Clinical Workflows and Precision Medicine'. *Journal of Biomedical Informatics* 108: 103479.	The central argument of this paper is that even if an AI-based clinical decision support system has been demonstrated to be accuracy and ethical, it will still fail to be adopted if clinicians cannot use it. Therefore, the authors argue that equal attention should be paid to the development of the interface as to the development of the model.
Terry, Amanda L. et al. 2022. 'Is Primary Health Care Ready for Artificial Intelligence? What Do Primary Health Care Stakeholders Say?' *BMC Medical Informatics and Decision Making* 22(1): 237.	This paper makes clear that healthcare practitioners have mixed views about the use of AI in frontline care – viewing its development and implementation as a 'double-edged sword.' The only guaranteed way of improving this perception, according to the authors, is to ensure that AI models are co-created with clinicians rather than just thrust upon them without consultation.
Upshaw, Tara L. et al. 2023. 'Priorities for Artificial Intelligence Applications in Primary Care: A Canadian Deliberative Dialogue with Patients, Providers, and Health System Leaders'. *The Journal of the American Board of Family Medicine* 36(2): 210–20.	This paper highlights the hopes (not yet realised) for AI., including the hope that AI will create more time and cognitive freedom for clinicians, enabling them to spend more time focusing on the social aspects of care, and helping to coordinate care and support for patients in the community.
Watson, Joshua et al. 2020. 'Overcoming Barriers to the Adoption and Implementation of Predictive Modeling and Machine Learning in Clinical Care: What Can We Learn from US Academic Medical Centers?' *JAMIA Open* 3(2): 167–72.	Based on observations of AI-tools in use, the authors highlight 4 best practices which all AI developers should abide by: (1) involve clinicians in the development of AI models; (2) design the model with a clear clinical intervention in mind; (3) identify and transparently report performance metrics; (4) regularly re-evaluate model performance.
Yoo, Junsang, Sujeong Hur, Wonil Hwang, and Won Chul Cha. 2023. 'Healthcare Professionals' Expectations of Medical Artificial Intelligence and Strategies for Its Clinical Implementation: A Qualitative Study'. *Healthcare Informatics Research* 29(1): 64–74.	This paper again highlights the mixed views of clinicians regarding the use of AI in clinical care. It notes that some healthcare practitioners do recognise the potential for AI to (e.g.,) improve patient safety, they are also concerned about distortions to workflow, automation bias, and alert fatigue.

Will patients and publics accept it?

Whilst considerable attention has been paid to patient and public attitudes regarding the use of health data for research purposes in general – in part because there have been so many public failures in this space[85] -, much less attention has been paid to patient and public attitudes regarding AI. What research has been done reveals a generally supportive attitude, despite a lack of knowledge about AI. In general, there seems to be a belief that the potential benefits of AI outweigh the risks, although this depends on whether individuals genuinely belief that AI is capable of improving the process of diagnosis and treatment management[86,87]. This is, however, neither a unilaterally held nor unconditional belief. Younger individuals, those with lower levels of educational attainment, and representatives from black and minority ethnic populations all show less belief in the potential benefits of AI[87] (potentially because individuals from within these groups have lower levels of trust in the healthcare system in general). In addition, some patients worry about 'uniqueness neglect' fearing that AI models may ignore their unique characteristics and circumstances, resulting in poorer quality care and worse outcomes[88], and others will only accept the use of AI in a purely supportive function[89]. Overall, it is clear that much more needs to be done to improve patient and public understanding of AI, what it is and how it works, and to improve public trust in the use of AI – for example, by being far more transparent about the development of AI policy and regulation[90].

Paper	**Why is it useful?**
Aggarwal, Ravi et al. 2021. 'Patient Perceptions on Data Sharing and Applying Artificial Intelligence to Health Care Data: Cross-Sectional Survey'. *Journal of Medical Internet Research* 23(8): e26162.	This paper reports the results of a large survey of patients from a London Hospital regarding their attitudes towards sharing health data for AI research. The study found that, despite a lack of knowledge about AI and ML, patients were more likely to be trusting of the idea of AI than not, and a large proportion thought that the potential benefits outweighed the potential risks. This was not an evenly held opinion, however, individuals from black and minority ethnic groups were less supportive of data sharing and AI, as well as younger patients and those with lower levels of educational attainment.
Carter, Pam, Graeme T. Laurie, and Mary Dixon-Woods. 2015. 'The Social Licence for Research: Why Care.Data Ran into Trouble'. *Journal of Medical Ethics* 41(5): 404–9.	This paper applies the concept of 'the social licence for research' to the case of Care.Data – a failed UK Government health data project. It makes clear that when it comes to public trust, and therefore public acceptance, of large data projects (which would now include those involving AI), compliance with the law is not sufficient.
Esmaeilzadeh, Pouyan. 2020. 'Use of AI-Based Tools for Healthcare Purposes: A Survey Study from Consumers' Perspectives'. *BMC Medical Informatics and Decision Making* 20(1): 170.	This paper reports the results of a study assessing consumer attitudes regarding the risks and benefits associated with the use of AI in clinical decision making. It shows that perceptions of risk and perceptions of benefit are inversely correlated: when the risks of AI are perceived to be high, then the benefits are perceived to be low (and vice versa). The authors argue, therefore, that to improve acceptance levels, more needs to be done to convince patient and publics of the potential for AI to improve diagnostics, prognosis, and patient management.
Longoni, Chiara, Andrea Bonezzi, and Carey K Morewedge. 2019. 'Resistance to Medical Artificial Intelligence'. *Journal of Consumer Research* 46(4): 629–50.	This paper hypothesises that consumers are resistant to the idea of using AI in healthcare because of their fear of uniqueness neglect i.e., the idea that AI models will ignore their unique circumstances and characteristics, resulting in poorer quality care and worse outcomes.
Mikkelsen, Josefine Graabaek et al. 2023. 'Patient Perspectives on Data Sharing Regarding Implementing and Using Artificial Intelligence in General Practice – a Qualitative Study'. *BMC Health Services Research* 23(1): 335.	This paper reports the results from an interview-based study assessing patient attitudes regarding the use of AI in primary care. The study found that patients were generally willing to share their health data for the purpose of AI development and use, provided any AI

	tool was only used as a support tool and GPs were still the primary decision maker.
Wu, Chenxi et al. 2023. 'Public Perceptions on the Application of Artificial Intelligence in Healthcare: A Qualitative Meta-Synthesis'. *BMJ Open* 13(1): e066322.	This meta-synthesis of public attitudes towards medical AI, find that while there is a general acknowledgement of the potential benefits of medical AI, there are also concerns regarding privacy, security, and regulation. The authors conclude, that to build public trust, more needs to be done to improve public understanding of AI and how it is regulated.

How is it Governed?

As has been made apparent, the successful deployment and adoption of AI in healthcare largely depends on the extent to which healthcare practitioners, patients, publics, and other stakeholders feel as though its development and use is sufficiently well governed[91,92]. Good governance is seen as being the key to protecting patient safety[93] and ensuring trustworthiness[83]. Governance systems are comprised of both 'hard' (regulatory and legal) and 'soft' (standards and policies) elements, and so it is necessary to consider both these elements in turn.

To start with 'hard Governance' the development, deployment, and use of AI for healthcare is legally complex. Medical device laws; anti-discrimination laws; medical negligence/liability laws; data protection laws; intellectual property laws; and consumer protection laws, are all relevant[94] and yet are all being disrupted by the development of AI.

- Any AI model or system that is intended for the purpose of displaying, analysing, or printing medical information about a patient, for the purpose of supporting or providing recommendations to a healthcare practitioner, for the purpose of enabling a healthcare practitioner to independently review the recommendations of another doctor or software application, is likely to be considered a medical device[95], yet evolving medical device laws in the UK, the US, and the EU all currently lack clarity and are fraught with contradictions – a fact that developers feel is a clear current barrier to innovation (particularly for SMEs)[96]. Of particular note, is the lack of clarity regarding the rules and responsibilities of 'post-market surveillance' when it comes to AI i.e., how should models be monitored after they are deployed? Do they need to be regularly re-evaluated? Does this depend on the type of model (e.g., adaptive or static? [97]
- The legality of medical data processing depends on context and intended purpose. If for example, data is being processed for research purposes then explicit and informed consent is usually required. If, however, data is being processed for purposes of 'direct care' or quality assurance purposes then consent is not usually required. AI models blend these different use cases: An AI system, based on a self-learning model, that is deployed inside a clinical system for the purposes of (a) providing diagnostic support to clinicians and (b) sending reports back to regulators regarding overall compliance rates with recommended treatment pathways is simultaneously being used for research, quality improvement, and direct care purposes[98]. Furthermore, even when consent is clearly required, the black box nature of some AI models makes it difficult to obtain consent in a way that is genuinely 'specific, informed, and unambiguous'[99,100].
- Negligence cases – both product and medical – depend on a claimant's ability to 'prove' causality of damages. In other words, a patient must be able to prove that they came to harm through the actions/inaction of a medical practitioner or using a faulty medical device (including software as a medical device)[101]. Identifying causality, and meeting the burden of proof, is however made exponentially more complicated by the number of actors and systems

involved in the development, deployment, and use of AI. Claims of negligence (and so liability) will likely depend on whether the clinician in question acted according to the medical standard, and whether or not the medical device was used as intended or in an 'off-label' manner. However, both concepts (medical standard and intended use) are made more fluid by the introduction of AI. What happens, for example, if the device 'adapts' beyond its intended use, but there is no requirement for it to be re-evaluated? Or, if the medical device provides advice that deviates from the medical standard but there is a policy in place stating that a clinician must always follow the advice of the algorithm? What happens if a patient claims that they were not able to give informed consent, because their treatment was determined by an algorithm that they cannot understand. All of these – and many other – questions currently remain open[102].

Moving to the 'soft' elements of Governance, the safe and trustworthy development, deployment, and use of AI also depends on the existence of policies and standards that ensure (as far as possible) fairness, transparency, and accountability[103] of AI models. Soft Governance is, therefore, a far broader (and more flexible/adaptable[104]) category than hard Governance and, consequently, covers a far wider range of topics. There is, however, a consensus developing around the core 'themes' of soft Governance – even if there is far less consensus regarding the operationalisation of each theme – including: registration of algorithms in use in clinical care[105]; open code[106]; transparent reporting of model training and evaluation datasets[107]; public recording of safety incidents involving AI[93]; and development of benchmarking datasets[108].

Paper – Hard Governance	**Why is it useful?**
Baines, Rebecca et al. 2022. 'Navigating Medical Device Certification: A Qualitative Exploration of Barriers and Enablers Amongst Innovators, Notified Bodies and Other Stakeholders'. *Therapeutic Innovation & Regulatory Science*.	This paper reports the findings of a qualitative study into AI developer attitudes regarding Medical Device Law. It concludes that whilst participants saw the necessity of medical device law, they described existing rules as unclear and subject to unnecessary and unhelpful political influence.
Blasimme, Alessandro, and Effy Vayena. 2020. 'The Ethics of AI in Biomedical Research, Patient Care, and Public Health'. In *The Oxford Handbook of Ethics of AI*, eds. Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press, 702–18.	This paper introduces a Governance Model that the authors term 'systemic oversight' based on six key concepts: adaptivity, flexibility, inclusiveness, reflexivity, responsiveness, and monitoring.
Butterworth, Michael. 2018. 'The ICO and Artificial Intelligence: The Role of Fairness in the GDPR Framework'. *Computer Law & Security Review* 34(2): 257–68.	This paper outlines how AI challenges the GDPR. It highlights how black box algorithms challenge the concept of meaningful and informed consent as well as 'the right of withdrawal.'
Cohen, I. Glenn et al. 2014. 'The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care'. *Health Affairs* 33(7): 1139–47.	This paper provides an incredibly detailed and useful overview of the many legal issues raised by AI. It covers both data protection laws and medical liability laws in particular detail, highlighting complexities related to intended data processing purpose and liabilities related to following/ignoring AI-based clinical recommendations.
Hwang, Thomas J., Aaron S. Kesselheim, and Kerstin N. Vokinger. 2019. 'Lifecycle Regulation of Artificial Intelligence– and Machine Learning–Based Software Devices in Medicine'. *JAMA* 322(23): 2285.	In this paper, the authors focus on the need for medical device regulation to cover the entire lifecycle of AI models - including 'post-deployment.' It notes that little is currently known about how to handle the fact that a model's performance might shift/drift over time and yet

	there is a need to put regulatory guardrails in place to safeguard patients from any potential harm that might occur as a result of this process.
Jackups, Ronald. 2023. 'FDA Regulation of Laboratory Clinical Decision Support Software: Is It a Medical Device?' *Clinical Chemistry* 69(4): 327–29.	This is a very up-to-date paper covering the development of medical device law in the US. It covers the definition of software as a medical device according to the FDA, but also critiques the latest guidance – noting that ambiguities and contradictions present within the guidance are currently causing developers considerable confusion and concern.
Molnár-Gábor, Fruzsina. 2020. 'Artificial Intelligence in Healthcare: Doctors, Patients and Liabilities'. In *Regulating Artificial Intelligence*, eds. Thomas Wischmeyer and Timo Rademacher. Cham: Springer International Publishing, 337–60.	This paper provides a phenomenally detailed discussion of the many different liability issues raised by the introduction of AI in healthcare. It offers a multitude of arguments and counterarguments and makes very clear that, yet, there are no straightforward answers to 'who is responsible/ should be held liable' if harm results from the use of an AI tool in the care of a patient.
Schönberger, D. 2019. 'Artificial Intelligence in Healthcare: A Critical Analysis of the Legal and Ethical Implications'. *International Journal of Law and Information Technology* 27(2): 171–203.	This paper provides a high-level overview of the legal complexities presented by the development of AI, focusing on data governance, medical device regulation, and intellectual property implications.
Smith, Helen, and Kit Fotheringham. 2022. 'Exploring Remedies for Defective Artificial Intelligence Aids in Clinical Decision-Making in Post-Brexit England and Wales'. *Medical Law International* 22(1): 33–51.	This paper acts like a legal thought experiment, reviewing different options and their likely outcomes for anyone wishing to bring a claim of negligence against a defective AI-based clinical decision support system in the UK in the wake of it leaving the EU and so leaving behind EU medical device and consumer protection laws.
Williams, Garrath, and Iris Pigeot. 2017. 'Consent and Confidentiality in the Light of Recent Demands for Data Sharing: Consent, Confidentiality, and Data Sharing'. *Biometrical Journal* 59(2): 240–50.	This paper offers readers a useful introduction to the many complexities and uncertainties regarding consent, accountability, and trustworthiness, in an age where the development of AI has significantly increased the demand for data inter-institutional and indeed international data sharing.

Paper – Soft Governance	**Why is it useful?**
Bedoya, Armando D et al. 2022. 'A Framework for the Oversight and Local Deployment of Safe and High-Quality Prediction Models'. *Journal of the American Medical Informatics Association* 29(9): 1631–36.	This paper makes a clear argument for the development of a Governance framework that covers the whole lifecycle of an AI model. Of note is its explicit plea for the introduction of a requirement that all algorithms be registered when in use.
Bozkurt, Selen et al. 2020. 'Reporting of Demographic Data and Representativeness in Machine Learning Models Using Electronic Health Records'. *Journal of the American Medical Informatics Association* 27(12): 1878–84.	The authors in this paper lament the fact that demographic descriptions of training data are currently poorly reported which makes it difficult to assess whether the training dataset was representative of the population upon which the AI model might be used. This has implications for the extent to which the model's fairness can be assessed. The authors suggest that a lot can be learned from the transparency requirements surrounding clinical trials, including preregistration, and making source-code openly available.
Char, Danton S., Michael D. Abràmoff, and Chris Feudtner. 2020. 'Identifying Ethical Considerations for	This paper provides a framework, covering all stages of AI model development, and can be used to help developers look ahead and identify issues that might

Machine Learning Healthcare Applications'. *The American Journal of Bioethics* 20(11): 7–17.	arise in the future or to ask questions in the moment. Specifically, the Framework is divided into three sections: (1) conception: auditability, transparency standards, and conflicts of interest; (2) calibration: accuracy, trading of test characteristics, and calibrated risk of harm; and (3) implementation, evaluation, and oversight, adverse events, ongoing assessment of accuracy and usage
Hernandez-Boussard, Tina, Selen Bozkurt, John P A Ioannidis, and Nigam H Shah. 2020. 'MINIMAR (MINimum Information for Medical AI Reporting): Developing Reporting Standards for Artificial Intelligence in Health Care'. *Journal of the American Medical Informatics Association* 27(12): 2011–15.	This paper too notes the importance of reporting the key information about the datasets used to train AI models. With the intention of improving transparency, the authors propose the Minimum Information for Medical AI Reporting (MINIMAR) standard covering four major reporting requirements.
Liao, Frank, Sabrina Adelaine, Majid Afshar, and Brian W. Patterson. 2022. 'Governance of Clinical AI Applications to Facilitate Safe and Equitable Deployment in a Large Health System: Key Elements and Early Successes'. *Frontiers in Digital Health* 4: 931439.	This paper provides a case study of the AI Governance structure in place at the University of Wisconsin Health facility. It provides a (rare) detailed example of AI Governance in action.
Macrae, Carl. 2019. 'Governing the Safety of Artificial Intelligence in Healthcare'. *BMJ Quality & Safety* 28(6): 495–98.	Focusing exclusively on the risks AI poses to patient safety, this paper outlines a number of preventative measures the author believes are necessary for mitigating risks. These include: the publication of safety reports, and the implementation of 'black box' recorders (like in airlines) to capture any data related to safety events.
Reddy, Sandeep, Sonia Allan, Simon Coghlan, and Paul Cooper. 2020. 'A Governance Model for the Application of AI in Health Care'. *Journal of the American Medical Informatics Association* 27(3): 491–97	This relatively detailed paper outlines a proposed Governance Model comprised of 4 main components: fairness, transparency, trustworthiness, and accountability.
Wiens, Jenna et al. 2019. 'Do No Harm: A Roadmap for Responsible Machine Learning for Health Care'. *Nature Medicine* 25(9): 1337–40.	This is a far more technical paper than the others in this category, focusing more on how requirements for representative datasets might be met in practice, including the use of synthetic data produced by generative adversarial networks.

What about do no harm?

Medicine, despite being one of the most highly regulated 'industries' in existence, is not purely governed by rules, regulations, policies, and standards. It also has a long (not always successful) history of ethical governance from the Hippocratic Oath to 'do no harm', to the introduction of the bioethics principles (autonomy, beneficence, non-maleficence, justice), and finally to more recent Medicine-adjacent ethical interventions such the Bermuda Principles intended to govern human genome sequencing. It is necessary, therefore, to ensure the introduction of AI is also subject to rigorous ethical analysis, alongside technical, regulatory, and sociocultural analysis. This can be done by first applying the expanded list of bioethics principles (autonomy, beneficence, non-maleficence, justice, Explainability) commonly used for the ethical analysis of AI in general, to the analysis of the ethics of AI for healthcare, and second by considering the broader value-based implications[109]:

- '**Autonomy**' broadly refers to the ability of a person to make their own life. It is a key concept in Western moral and political philosophy and is protected/harmed by a person's ability/inability to self-govern in a manner that is free from external control and undue interference[110]. To a large extent, in modern medicine, autonomy is seen as being the 'primary principle' and the need to protect autonomy has been attached to several significant shifts in modern medicine, including the shift from 'paternalistic care' towards 'patient-centred care'[111]. AI's reliance on large volumes of data, including data that patients may collect themselves (e.g., smartwatches or shopping records), and its ability to predict risk – with the intention of encouraging preventative action – means that, without careful thought, AI has significant potential to nudge and police individuals into behaving in ways that do not necessarily align with their own personal values in the name of pursuing 'optimum health.' The fact that much of this algorithmic nudging happens within a black box, and involves the evaluation of an individual against often inscrutable baselines amplifies the potential for AI to have a negative impact on autonomy[112]. Other considerations, include the fact that autonomy is primarily protected by an individual's right to informed and meaningful consent (which has already highlighted is not always upheld in the context of AI) and their right to 'not know' if they think certain health information (such as that involving future risk)[113] might cause them psychological harm[114], both of which are potentially disrupted by the potential for AI to usher in an age of near continuous unobservable screening process[115]. These reasons, and others, highlight why arguments that centre on the idea that AI will be empowering for individuals are flawed[116].

- **'Beneficence'** broadly refers to the duty of healthcare providers to both prevent/remove harm and to promote wellbeing/ welfare. This involves more than 'just' identifying a diagnosis that fits a list of quantifiable symptoms and matching this to an 'effective' drug, or identifying potential risk factors, beneficent care also involves seeing the person as a whole (taking into account their personal beliefs, values, etc.), shared decision-making, and providing care in a manner in an empathetic, compassionate, and trustworthy manner[117]. Whilst AI might be able to mimic empathy it cannot truly 'understand' it and therefore might not be able to completely replicate its effects. Furthermore, there is a growing concern that AI's reliance on 'quantifiable' data might lead to the exclusion of other 'data' about a patient's life in the decision-making process[118,119]. For these reasons, it is important to see AI as a helpful aide, but not a replacement for clinicians who are capable of contextualising 'evidence' and focusing on the 'softer' aspects of care[120].
- **'Non-Maleficence'** is the principle most closely linked to the Hippocratic 'Do No Harm' oath. Here, the concerns raised by AI mostly stem from its ability to do more harm than good, by (for example) infringing on patient privacy[121],; or leading to widespread 'overdiagnosis' that can cause both physical and psychological harm as well as result in waste[122–124]; or by enabling healthcare (for example, the definition of illness) to be unethically manipulated by economic and market forces[125].
- **'Justice'** is the most familiar of the ethical principles in the context of AI, at least in the public domain, given its close tie with issues of bias. The ethical concerns here are that issues with the way medical data (used to train AI) are collected, curated, and interpreted may lead to biased algorithms which, over time, might lead to discrimination [126,127]. Whilst most of the focus in the literature, and in the press, in this domain has been on the potential for AI to be biased in terms of sex, gender, or race[128], there are also lesser-known bias problems. Examples: include the potential for precision medicine (enabled by AI) to divide the population into 'good patients' deserving of care (those who respond well to treatments and act on preventive advice) and 'bad patients' undeserving of care (those who do not respond well to treatments and may be unable to act on preventive advice[129]; latent bias (i.e., bias that develops over time)[130]; and the potential for AI to amplify the effects of the inverse care law (i.e., those who are in greatest need of care are least able to access it)[131]. If the widespread implementation of AI is to be 'successful',

then it will be necessary to question any assumptions that algorithms are somehow more objective[132], and to develop a range of mechanisms for dealing with the sources of bias, and for identifying the consequences[133,134].

- **'Explainability'** – the ability to 'explain' or understand how an algorithm reaches a 'decision – is, in many ways, an umbrella principle with its importance being justified for purposes linked to all preceding principles. For example, there are legal (autonomy/justice) justifications linked to the value of informed consent; and medical (beneficence/non-maleficence) justifications linked to the importance of detecting errors (e.g., incidents of spurious correlation being confused with causality) that might lead to direct harm via misdiagnosed or missed diagnosis, or indirect harm via overdiagnosis[135]. It is perhaps because of this overarching principle that Explainability is the principle that has been 'operationalised' most successfully via efforts of the XAI community[136].

Paper – Autonomy	**Why is it useful?**
Andorno, R. 2004. 'The Right Not to Know: An Autonomy Based Approach'. *Journal of Medical Ethics* 30(5): 435–39.	This paper is not specific to AI, but its argument can be helpfully extrapolated. Taking a meta-ethics approach, the author argues that just as a patient has a right to information regarding their health, they also have the right to give up this right i.e., they have a right not to know certain information about their health should they finding it (for example) distressing. Such information might, for instance, relate to the likelihood of a person developing a specific disesase (risk prediction). As many AI systems will effectively act as 'always on' screening tools, this right not to know might be subverted potentially unwittingly undermining patient autonomy.
Blasimme, Alessandro, and Effy Vayena. 2016. 'Becoming Partners, Retaining Autonomy: Ethical Considerations on the Development of Precision Medicine'. *BMC Medical Ethics* 17(1): 67.	This paper provides a detailed discussion of the concept of autonomy and how it might be impacted by the development of precision medicine (which as discussed is enabled by AI). The authors take a moral philosophical approach to discussing the concept and explain how precision medicine's need for ever increasing volumes of detailed data lures people into becoming complicit and participants in their own bodily surveillance which, again, may negatively impact their autonomy.
Green, S, and H Vogt. 2016. 'Personalizing Medicine: Disease Prevention in Silico and in Socio.' *HUMANA MENTE Journal of Philosophical Studies* 9(30): 105–45.	This paper explains the connection between AI, P4 medicine and continual screening, expounding on the argument that this can act as a form of seemingly 'beneficent control' that may have negative consequences for patient autonomy.
Grote, Thomas, and Philipp Berens. 2020. 'On the Ethics of Algorithmic Decision-Making in Healthcare'. *Journal of Medical Ethics* 46(3): 205–11.	In this paper the authors outline the many reasons why the introduction of AI into medical care might risk the reintroduction of a paternalistic approach to medical decision making. It focuses on the fact that algorithms may 'rank' potential treatment options according to purely quantitative measures, not accounting for a patient's preferences regarding values or quality of life requirements, and yet both patient and doctor might feel pressured to accept the algorithm's first recommended option. This would undermine the shared decision-making model that has become the hallmark of autonomy-supporting medical care in recent years.
Hofmann, Bjørn, and Michal Stanak. 2018. 'Nudging in Screening: Literature Review and Ethical Guidance'. *Patient Education and Counseling* 101(9): 1561–69.	This paper may also not appear directly relevant at first. However, it focuses on the argument that screening programmes can be operated in a way that is

	overly paternalistic, undermines free choice, and disrupts shared decision making. – all factors essential to the support of patient autonomy. Again, as AI is likely to be implemented as a 'screening tool' these issues of nudging or manipulating people into certain actions, these ethical threats to patient autonomy also apply to AI.
Morley, Jessica, and Luciano Floridi. 2020. 'The Limits of Empowerment: How to Reframe the Role of MHealth Tools in the Healthcare Ecosystem'. *Science and Engineering Ethics* 26(3): 1159–83.	In this paper, we analyse the limitations of the empowerment narrative (i.e., the idea that fiving patients and consumers access to more and more personalised data will automatically empower them to take better care of their own health) from a variety of philosophical, moral, and ethical perspectives.
Schwartz, Peter H., and Eric M. Meslin. 2008. 'The Ethics of Information: Absolute Risk Reduction and Patient Understanding of Screening'. *Journal of General Internal Medicine* 23(6): 867–70.	This paper makes clear the potential for the presentation of 'absolute' or even 'relative' risk information to cause psychological harm if not handled correctly and cautiously, with an understanding of the specific patient's context and in particular their level of digital health literacy.

Paper – Beneficence	**Why is it useful?**
Chin-Yee, Benjamin, and Ross Upshur. 2019. 'Three Problems with Big Data and Artificial Intelligence in Medicine'. *Perspectives in Biology and Medicine* 62(2): 237–56.	This paper provides a useful high-level introduction to the argument that big data and its use in medicine (for example via AI) might usher in a return to logical positivism in medicine – i.e., the idea that what is observed, what is represented in data, is 'fact' and not the result of social processes over which the patient may have little or no control.
Heyen, Nils B., and Sabine Salloch. 2021. 'The Ethics of Machine Learning-Based Clinical Decision Support: An Analysis through the Lens of Professionalisation Theory'. *BMC Medical Ethics* 22(1): 112.	This paper examines the evolving role of the clinician in an increasingly data-driven or algorithmically enhanced healthcare system. It highlights the need for human clinicians to focus on the 'soft' facts of a patient's case that cannot be comprehended by an AI model e.g., the patient's personality, life situation or cultural or religious background.
Kerasidou, Angeliki. 2020. 'Artificial Intelligence and the Ongoing Need for Empathy, Compassion and Trust in Healthcare'. *Bulletin of the World Health Organization* 98(4): 245–50.	This paper too notes the continuing need for human care despite the rapid evolution of AI, noting that empathy, compassion, and trust are fundamental values of patient-centred care and whilst AI might promise greater efficiency, and effectiveness, and a level of personalisation not possible before, it will not be able to replicate the essential and necessary human aspects of care.
McDougall, Rosalind J. 2019. 'Computer Knows Best? The Need for Value-Flexibility in Medical AI'. *Journal of Medical Ethics* 45(3): 156–60.	This paper convincingly argues the importance of considering personal and societal values in medical decision making and the importance of recognising that these might differ depending on a variety of sociocultural factors. The authors state that human clinicians (although not always) are capable of 'value flexibility' and so adapting their decisions to suit a particular patient's values, and that the positive impact this ability has on patient outcomes suggests that AI models must be designed with this capacity too.

Paper – Non-Maleficence	**Why is it useful?**
Bartoletti, Ivana. 2019. 'AI in Healthcare: Ethical and Privacy Challenges'. In *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, eds. David Riaño,	This paper acts as a high-level primer on the ethical issues raised by potential privacy infringements that might result from the development, deployment, and use of AI.

Szymon Wilk, and Annette ten Teije. Cham: Springer International Publishing, 7–10.	
Fritzsche, Marie-Christine et al. 2023. 'Ethical Layering in AI-Driven Polygenic Risk Scores—New Complexities, New Challenges'. *Frontiers in Genetics* 14: 1098439.	By focusing on complex risk scores, this paper highlights the ethical issues raised by the potential for spurious correlations to be (mis)interpreted as evidence of causality, and the implications this has for 'deterministic' attitudes to develop.
McCartney, Margaret et al. 2020. 'Why "Case Finding" Is Bad Science'. *Journal of the Royal Society of Medicine* 113(2): 54–58.	Based on an analysis of the principles usually used to guide decisions about the introduction of screening processes, this paper argues that introducing ubiquitous screening programmes (for example via AI) can result in inequity and bypass the long-established safety inherent in scrutiny and governance from the organisations designed to protect the public from non-evidence-based screening programmes.
Rubeis, Giovanni. 2023. 'Liquid Health. Medicine in the Age of Surveillance Capitalism'. *Social Science & Medicine* 322: 115810.	This paper makes clear the implications of separating knowledge about the body and knowledge about medicine from patients and from the relatively 'walled garden' of the clinical community and instead placing this knowledge in algorithms, monitoring devices, and private companies. The author explains how this could result in economic manipulation of healthcare.
Vogt, Henrik, Sara Green, Claus Thorn Ekstrøm, and John Brodersen. 2019. 'How Precision Medicine and Screening with Big Data Could Increase Overdiagnosis'. *BMJ*: l5270.	Picking up on some of the themes also raised by McCartney et al (above), the authors in this paper argue that AI-based screening and risk prediction might result in an 'overdiagnosis' problem which might be both wasteful and directly harmful.

Paper – Justice	**Why is it useful?**
Abettan, Camille. 2016. 'Between Hype and Hope: What Is Really at Stake with Personalized Medicine?' *Medicine, Health Care and Philosophy* 19(3): 423–30.	This paper elucidates the argument that personalised medicine (which is closely linked to AI) to increase the risk of discrimination not necessarily based on known demographic sources of bias, but also the potential for personalised medicine to divide the population into responders (i.e. good patients deserving of care) and non-responders (i.e., bad patients underserving care) to the (e.g.,) preventative advice provided by AI models. Such clear-cut divisions fail to account for socioeconomic factors that might, for instance prevent some people from taking advantage of preventive advice. It is a nuanced and highly detailed argument.
Avellan, Tero, Sumita Sharma, and Markku Turunen. 2020. 'AI for All: Defining the What, Why, and How of Inclusive AI'. In *Proceedings of the 23rd International Conference on Academic Mindtrek*, AcademicMindtrek '20, New York, NY, USA: Association for Computing Machinery, 142–44.	Rather than purely focusing on the sources and potential consequences of bias in AI, this paper also makes several recommendations to counter these risks and make AI as inclusive as possible. These recommendations centre around three core ideas: the need for algorithms to be designed by diverse teams; the need for training data to representative; and the need for AI to be accessible to all users.
Cirillo, Davide et al. 2020. 'Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare'. *npj Digital Medicine* 3(1): 81.	This paper makes an important point that is often missed in purely technical (i.e., non-medical) discussions of bias and healthcare. The authors stress that whilst discriminatory bias is problematic, can lead to inequity in care, and should be actively countered, there are some situations in medical care in which bias is desirable for example when accounting for factors such as gender, sex, race might be necessary to achieve a more precise or accurate diagnosis.
DeCamp, Matthew, and Charlotta Lindvall. 2020. 'Latent Bias and the Implementation of Artificial	Whilst most papers focus on the 'upfront' sources of bias e.g., bias in the data used to train models, this

Intelligence in Medicine'. *Journal of the American Medical Informatics Association* 27(12): 2020–23.	paper focuses on 'latent bias' i.e., bias that might develop overtime from algorithmic or population drift, from error, or from other contextual or technical factors.
Gianfrancesco, Milena A., Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. 'Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data'. *JAMA Internal Medicine* 178(11): 1544.	This is a relatively technical paper discussing the potential issues in EHR data that might lead to bias in AI models. In particular it notes that data is often missing in a non-random fashion from EHR data (e.g., ethnicity is less likely to be recorded for some communities), that some diseases have naturally small sample sizes, and that EHR data might include errors such as misclassification of disease that a model might learn and these errors might be more likely to occur in the examination and treatment of some patients than in others.
Gray, Muir, Tyra Lagerberg, and Viktor Dombrádi. 2017. 'Equity and Value in "Precision Medicine"'. *The New Bioethics* 23(1): 87–94.	This paper makes the connection between the Inverse Care Law – i.e., the fact that those most in need of care are least likely to be able to access it – and AI. For example, the authors note that wealthier individuals (and often 'healthier individuals) are more likely to have devices that self-track, or to be able to pay for genome sequencing, thus generating more information on themselves and enabling the development of more accurate 'personalised algorithms' creating a self-reinforcing feedback loop of benefit.
McCradden, Melissa D et al. 2020. 'Patient Safety and Quality Improvement: Ethical Principles for a Regulatory Approach to Bias in Healthcare Machine Learning'. *Journal of the American Medical Informatics Association* 27(12): 2024–27.	The authors in this paper note that bias is not an issue that is exclusive to AI, human clinicians are also often biased, but that the introduction of AI does raise the stakes – for example, by increasing the potential scale of harm, or hiding sources of bias within the veneer of 'algorithmic objectivity' making it harder to spot and correct.
Parikh, Ravi B., Stephanie Teeple, and Amol S. Navathe. 2019. 'Addressing Bias in Artificial Intelligence in Health Care'. *JAMA* 322(24): 2377.	This is a detailed technical paper that discusses the causes and corrective measures of statistical bias and contrasts this to the causes and corrective measures of social bias. It highlights the fact that to help ensure the 'fair' application of AI to healthcare – both sources and types of bias need to be paid equal levels of attention.
Paulus, Jessica K., and David M. Kent. 2020. 'Predictably Unequal: Understanding and Addressing Concerns That Algorithmic Clinical Prediction May Increase Health Disparities'. *npj Digital Medicine* 3(1): 99.	This paper describes the various statistical measures available for testing algorithms for bias, providing a detailed overview of their various strengths and limitations. It concludes with a framework that can be used to test AI models for unfairness and bias depending on different types of predictions.
Verheij, Robert A, Vasa Curcin, Brendan C Delaney, and Mark M McGilchrist. 2018. 'Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse'. *Journal of Medical Internet Research* 20(5): e185.	This is the 'start at the very beginning paper' and outlines the many ways in which EHR data might become biased that are unrelated to demographic characteristics of patients. Specifically, it notes the following four potential sources of bias: delivery of care (there must be an event that can be recorded; recorded in EHR (an event that is not recorded will not be present in any dataset); extraction from EHR (data must be extracted for further analysis or reporting); and translation into database (extracted data must be re-databased as preparation for further analysis or reporting).

Paper – Explainability	**Why is it useful?**
Amann, Julia et al. 2020. 'Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary	This is a detailed explanation of the importance of Explainability, making the case from technological,

Perspective'. *BMC Medical Informatics and Decision Making* 20(1): 310.	legal, medical, and patient perspectives. For example, from a legal perspective, it raises the issues of informed consent, medical device laws, and liability; and from the medical perspective it highlights the need for clinicians to be able to question the conclusions reached by AI models so that they can identify any potential errors before they cause harm.
Price, W. Nicholson. 2018. 'Big Data and Black-Box Medical Algorithms'. *Science Translational Medicine* 10(471): eaao5333.	This is a reasonably technical, but extremely useful paper that introduces the different types of algorithms – how some are more 'black box' than others, and notes that black box algorithms should not automatically be dismissed because there are trade-offs to be made – for example, black box models might be more accurate than non-black box or they might have higher degrees of specificity.

Conclusion

Whilst undoubtedly there will be gaps in this guide, I hope that it has made clear the many complexities surrounding the development, deployment, and use of AI in healthcare, and it will help guide thoughtful and considered conversations as we move towards algorithmically-enhanced healthcare.

Bibliography

1. Reisman, Y. Computer-based clinical decision aids. A review of methods and assessment of systems. *Med. Inform. (Lond.)* **21**, 179–197 (1996).
2. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
3. Nwanosike, E. M., Conway, B. R., Merchant, H. A. & Hasan, S. S. Potential applications and performance of machine learning techniques and algorithms in clinical practice: A systematic review. *Int. J. Med. Inf.* **159**, 104679 (2022).
4. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* **19**, 1236–1246 (2018).
5. Almeida, J. R., Figueira Silva, J., Pazos, A., Matos, S. & Oliveira, J. L. Enhancing Decision-making Systems with Relevant Patient Information by Leveraging Clinical Notes. in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5 HEALTHINF: HEALTHINF,* 254–262 (Science and Technology Publications, 2020). doi:10.5220/0009166902540262.
6. Ellahham, S., Ellahham, N. & Simsekler, M. C. E. Application of Artificial Intelligence in the Health Care Safety Context: Opportunities and Challenges. *Am. J. Med. Qual.* **35**, 341–348 (2020).
7. Hall, P. S. & Morris, A. Predictive Analytics and Population Health. in *Key Advances in Clinical Informatics* 217–225 (Elsevier, 2017). doi:10.1016/B978-0-12-809523-2.00015-7.
8. Leyens, L., Reumann, M., Malats, N. & Brand, A. Use of big data for drug development and for public and personal health and care: Leyens et al. *Genet. Epidemiol.* **41**, 51–60 (2017).
9. Albahri, A. S. *et al.* A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* **96**, 156–191 (2023).
10. Will ChatGPT transform healthcare? *Nat. Med.* **29**, 505–506 (2023).
11. Homolak, J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croat. Med. J.* **64**, 1–3 (2023).
12. Castaneda, C. *et al.* Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *J. Clin. Bioinforma.* **5**, 4 (2015).
13. Blaser, R. *et al.* Improving pathway compliance and clinician performance by using information technology. *Int. J. Med. Inf.* **76**, 151–156 (2007).
14. Shapiro, D. W., Lasker, R. D., Bindman, A. B. & Lee, P. R. Containing Costs While Improving Quality of Care: The Role of Profiling and Practice Guidelines. *Annu. Rev. Public Health* **14**, 219–241 (1993).
15. Ciapponi, A. *et al.* Reducing medication errors for adults in hospital settings. *Cochrane Database Syst. Rev.* **11**, CD009985 (2021).
16. Ahmed, Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Hum. Genomics* **14**, 35 (2020).
17. Bousquet, J. *et al.* Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med.* **3**, 43 (2011).
18. Bulgarelli, L., Deliberato, R. O. & Johnson, A. E. W. Prediction on critically ill patients: The role of “big data”. *J. Crit. Care* **60**, 64–68 (2020).
19. Saqi, M. *et al.* Systems Medicine: The Future of Medical Genomics, Healthcare, and Wellness. in *Systems Medicine* (eds. Schmitz, U. & Wolkenhauer, O.) vol. 1386 43–60 (Springer New York, 2016).
20. Savard, J. Personalised Medicine: A Critique on the Future of Health Care. *J. Bioethical Inq.* **10**, 197–203 (2013).
21. Wyatt, D., Lampon, S. & McKevitt, C. Delivering healthcare’s ‘triple aim’: electronic health records and the health research participant in the UK National Health Service. *Sociol. Health Illn.* **42**, 1312–1327 (2020).
22. Deeny, S. R. & Steventon, A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Qual. Saf.* **24**, 505–515 (2015).
23. Obermeyer, Z. & Emanuel, E. J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).
24. Cosgriff, C. V., Celi, L. A. & Stone, D. J. Critical Care, Critical Data. *Biomed. Eng. Comput. Biol.* **10**, 117959721985656 (2019).
25. Sendak, M. *et al.* A Path for Translation of Machine Learning Products into Healthcare Delivery. *EMJ Innov.* (2020) doi:10.33590/emjinnov/19-00172.
26. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **20**, e262–e273 (2019).
27. De Silva, D. & Alahakoon, D. An artificial intelligence life cycle: From conception to production. *Patterns* **3**, 100489 (2022).
28. Chen, P.-H. C., Liu, Y. & Peng, L. How to develop machine learning models for healthcare. *Nat. Mater.* **18**, 410–414 (2019).
29. Muehlematter, U. J., Daniore, P. & Vokinger, K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit. Health* **3**, e195–e203 (2021).

30. Awaysheh, A. *et al.* Review of Medical Decision Support and Machine-Learning Methods. *Vet. Pathol.* **56**, 512–525 (2019).
31. Baalen, S., Boon, M. & Verhoef, P. From clinical decision support to clinical reasoning support systems. *J. Eval. Clin. Pract.* **27**, 520–528 (2021).
32. Zikos, D. A Framework to Design Successful Clinical Decision Support Systems. in *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments* 185–188 (ACM, 2017). doi:10.1145/3056540.3064960.
33. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1419–1428 (2018).
34. Osop, H. & Sahama, T. Systems Design Framework for a Practice-Based Evidence Approached Clinical Decision Support Systems. in *Proceedings of the Australasian Computer Science Week Multiconference* 1–6 (ACM, 2019). doi:10.1145/3290688.3290742.
35. Prausnitz, S., Altschuler, A., Herrinton, L. J., Avins, A. L. & Corley, D. A. The implementation checklist: A pragmatic instrument for accelerating RESEARCH-TO-IMPLEMENTATION cycles. *Learn. Health Syst.* (2023) doi:10.1002/lrh2.10359.
36. de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digit. Med.* **5**, 2 (2022).
37. Lisboa, P. J. G. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw.* **15**, 11–39 (2002).
38. Mahadevaiah, G. *et al.* Artificial intelligence‐based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Med. Phys.* **47**, (2020).
39. Miller, P. L. The evaluation of artificial intelligence systems in medicine. *Comput. Methods Programs Biomed.* **22**, 3–11 (1986).
40. Nsoesie, E. O. Evaluating Artificial Intelligence Applications in Clinical Settings. *JAMA Netw. Open* **1**, e182658 (2018).
41. Park, S. H. & Han, K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* **286**, 800–809 (2018).
42. Ayers, J. W. *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern. Med.* (2023) doi:10.1001/jamainternmed.2023.1838.
43. Gilson, A. *et al.* How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **9**, e45312 (2023).
44. England, J. R. & Cheng, P. M. Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. *Am. J. Roentgenol.* **212**, 513–519 (2019).
45. Neves, M. R. & Marsh, D. W. R. Modelling the Impact of AI for Clinical Decision Support. in *Artificial Intelligence in Medicine* (eds. Riaño, D., Wilk, S. & ten Teije, A.) vol. 11526 292–297 (Springer International Publishing, 2019).
46. Doyal, L. Need for moral audit in evaluating quality in health care. *Qual. Saf. Health Care* **1**, 178–183 (1992).
47. Liu, V. X., Bates, D. W., Wiens, J. & Shah, N. H. The number needed to benefit: estimating the value of predictive analytics in healthcare. *J. Am. Med. Inform. Assoc.* **26**, 1655–1659 (2019).
48. Kwan, J. L. *et al.* Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ* m3216 (2020) doi:10.1136/bmj.m3216.
49. Yusuf, M. *et al.* Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open* **10**, e034568 (2020).
50. Coiera, E. & Tong, H. Replication studies in the clinical decision support literature-frequency, fidelity, and impact. *J. Am. Med. Inform. Assoc.* **28**, 1815–1825 (2021).
51. Ge, W., Lueck, C., Suominen, H. & Apthorp, D. Has machine learning over-promised in healthcare? *Artif. Intell. Med.* **139**, 102524 (2023).
52. Andaur Navarro, C. L. *et al.* Systematic review finds "Spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J. Clin. Epidemiol.* S0895435623000756 (2023) doi:10.1016/j.jclinepi.2023.03.024.
53. Plana, D. *et al.* Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Netw. Open* **5**, e2233946 (2022).
54. Vasey, B. *et al.* Association of Clinician Diagnostic Performance With Machine Learning-Based Decision Support Systems: A Systematic Review. *JAMA Netw. Open* **4**, e211276 (2021).
55. Panch, T., Mattie, H. & Celi, L. A. The "inconvenient truth" about AI in healthcare. *Npj Digit. Med.* **2**, 77 (2019).
56. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).

57. Bainbridge, M. Big Data Challenges for Clinical and Precision Medicine. in *Big Data, Big Challenges: A Healthcare Perspective* (eds. Househ, M., Kushniruk, A. W. & Borycki, E. M.) 17–31 (Springer International Publishing, 2019). doi:10.1007/978-3-030-06109-8_2.
58. Dhindsa, K., Bhandari, M. & Sonnadara, R. R. What's holding up the big data revolution in healthcare? *BMJ* k5357 (2018) doi:10.1136/bmj.k5357.
59. Baxter, S. L. & Lee, A. Y. Gaps in standards for integrating artificial intelligence technologies into ophthalmic practice. *Curr. Opin. Ophthalmol.* **32**, 431–438 (2021).
60. Kerasidou, C. (Xaroula), Malone, M., Daly, A. & Tava, F. Machine learning models, trusted research environments and UK health data: ensuring a safe and beneficial future for AI development in healthcare. *J. Med. Ethics* jme-2022-108696 (2023) doi:10.1136/jme-2022-108696.
61. Lavin, A. *et al.* Technology readiness levels for machine learning systems. *Nat. Commun.* **13**, 6039 (2022).
62. Bucur, A. *et al.* Workflow-driven clinical decision support for personalized oncology. *BMC Med. Inform. Decis. Mak.* **16**, 87 (2016).
63. Goff, M. *et al.* Ambiguous workarounds in policy piloting in the NHS: Tensions, trade-offs and legacies of organisational change projects. *New Technol. Work Employ.* **36**, 17–43 (2021).
64. Rezaei-Yazdi, A. & Buckingham, C. D. Capturing Human Intelligence for Modelling Cognitive-Based Clinical Decision Support Agents. in *Artificial Life and Intelligent Agents* (eds. Lewis, P. R., Headleand, C. J., Battle, S. & Ritsos, P. D.) 105–116 (Springer International Publishing, 2018). doi:10.1007/978-3-319-90418-4_9.
65. Heckman, G. A., Hirdes, J. P. & McKelvie, R. S. The Role of Physicians in the Era of Big Data. *Can. J. Cardiol.* **36**, 19–21 (2020).
66. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: Empirical results assessing influencing factors. *Int. J. Med. Inf.* **83**, 368–375 (2014).
67. Buchlak, Q. D. *et al.* Ethical thinking machines in surgery and the requirement for clinical leadership. *Am. J. Surg.* **220**, 1372–1374 (2020).
68. Dullabh, P. *et al.* The Technology Landscape of Patient-Centered Clinical Decision Support – Where Are We and What Is Needed? in *Studies in Health Technology and Informatics* (eds. Otero, P., Scott, P., Martin, S. Z. & Huesing, E.) (IOS Press, 2022). doi:10.3233/SHTI220094.
69. Fosch-Villaronga, E., Drukarch, H., Khanna, P., Verhoef, T. & Custers, B. Accounting for diversity in AI for medicine. *Comput. Law Secur. Rev.* **47**, 105735 (2022).
70. Scobie, S. & Castle-Clarke, S. Implementing learning health systems in the UK NHS: Policy actions to improve collaboration and transparency and support innovation and better use of analytics. *Learn. Health Syst.* **4**, (2020).
71. Fridsma, D. B. Health informatics: a required skill for 21st century clinicians. *BMJ* **362**, k3043 (2018).
72. Yoo, J., Hur, S., Hwang, W. & Cha, W. C. Healthcare Professionals' Expectations of Medical Artificial Intelligence and Strategies for its Clinical Implementation: A Qualitative Study. *Healthc. Inform. Res.* **29**, 64–74 (2023).
73. Upshaw, T. L. *et al.* Priorities for Artificial Intelligence Applications in Primary Care: A Canadian Deliberative Dialogue with Patients, Providers, and Health System Leaders. *J. Am. Board Fam. Med.* **36**, 210–220 (2023).
74. Nitiéma, P. Artificial Intelligence in Medicine: Text Mining of Health Care Workers' Opinions. *J. Med. Internet Res.* **25**, e41138 (2023).
75. Terry, A. L. *et al.* Is primary health care ready for artificial intelligence? What do primary health care stakeholders say? *BMC Med. Inform. Decis. Mak.* **22**, 237 (2022).
76. Abouzahra, M. & Guenter, D. Exploring physicians' continuous use of clinical decision support systems. *Eur. J. Inf. Syst.* 1–22 (2022) doi:10.1080/0960085X.2022.2119172.
77. Watson, J. *et al.* Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* **3**, 167–172 (2020).
78. Rundo, L., Pirrone, R., Vitabile, S., Sala, E. & Gambino, O. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *J. Biomed. Inform.* **108**, 103479 (2020).
79. Braun, M., Hummel, P., Beck, S. & Dabrock, P. Primer on an ethics of AI-based decision support systems in the clinic. *J. Med. Ethics* **47**, e3–e3 (2021).
80. Jones, C., Thornton, J. & Wyatt, J. C. Enhancing trust in clinical decision support systems: a framework for developers. *BMJ Health Care Inform.* **28**, e100247 (2021).
81. Choudhury, A. Factors influencing clinicians' willingness to use an AI-based clinical decision support system. *Front. Digit. Health* **4**, 920662 (2022).
82. Catchpole, K. & Alfred, M. Industrial Conceptualization of Health Care Versus the Naturalistic Decision-Making Paradigm: Work as Imagined Versus Work as Done. *J. Cogn. Eng. Decis. Mak.* **12**, 222–226 (2018).
83. Crigger, E. *et al.* Trustworthy Augmented Intelligence in Health Care. *J. Med. Syst.* **46**, 12 (2022).
84. Kealey, E., Leckman-Westin, E. & Finnerty, M. T. Impact of four training conditions on physician use of a web-based clinical decision support system. *Artif. Intell. Med.* **59**, 39–44 (2013).

85. Carter, P., Laurie, G. T. & Dixon-Woods, M. The social licence for research: why *care.data* ran into trouble. *J. Med. Ethics* **41**, 404–409 (2015).
86. Esmaeilzadeh, P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC Med. Inform. Decis. Mak.* **20**, 170 (2020).
87. Aggarwal, R., Farag, S., Martin, G., Ashrafian, H. & Darzi, A. Patient Perceptions on Data Sharing and Applying Artificial Intelligence to Health Care Data: Cross-sectional Survey. *J. Med. Internet Res.* **23**, e26162 (2021).
88. Longoni, C., Bonezzi, A. & Morewedge, C. K. Resistance to Medical Artificial Intelligence. *J. Consum. Res.* **46**, 629–650 (2019).
89. Mikkelsen, J. G., Sørensen, N. L., Merrild, C. H., Jensen, M. B. & Thomsen, J. L. Patient perspectives on data sharing regarding implementing and using artificial intelligence in general practice – a qualitative study. *BMC Health Serv. Res.* **23**, 335 (2023).
90. Wu, C. *et al.* Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. *BMJ Open* **13**, e066322 (2023).
91. Liao, F., Adelaine, S., Afshar, M. & Patterson, B. W. Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes. *Front. Digit. Health* **4**, 931439 (2022).
92. Char, D. S., Abràmoff, M. D. & Feudtner, C. Identifying Ethical Considerations for Machine Learning Healthcare Applications. *Am. J. Bioeth.* **20**, 7–17 (2020).
93. Macrae, C. Governing the safety of artificial intelligence in healthcare. *BMJ Qual. Saf.* **28**, 495–498 (2019).
94. Schönberger, D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *Int. J. Law Inf. Technol.* **27**, 171–203 (2019).
95. Jackups, R. FDA Regulation of Laboratory Clinical Decision Support Software: Is It a Medical Device? *Clin. Chem.* **69**, 327–329 (2023).
96. Baines, R. *et al.* Navigating Medical Device Certification: A Qualitative Exploration of Barriers and Enablers Amongst Innovators, Notified Bodies and Other Stakeholders. *Ther. Innov. Regul. Sci.* (2022) doi:10.1007/s43441-022-00463-4.
97. Hwang, T. J., Kesselheim, A. S. & Vokinger, K. N. Lifecycle Regulation of Artificial Intelligence– and Machine Learning–Based Software Devices in Medicine. *JAMA* **322**, 2285 (2019).
98. Cohen, I. G., Amarasingham, R., Shah, A., Xie, B. & Lo, B. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. *Health Aff. (Millwood)* **33**, 1139–1147 (2014).
99. Butterworth, M. The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Comput. Law Secur. Rev.* **34**, 257–268 (2018).
100. Williams, G. & Pigeot, I. Consent and confidentiality in the light of recent demands for data sharing: Consent, confidentiality, and data sharing. *Biom. J.* **59**, 240–250 (2017).
101. Smith, H. & Fotheringham, K. Exploring remedies for defective artificial intelligence aids in clinical decision-making in post-Brexit England and Wales. *Med. Law Int.* **22**, 33–51 (2022).
102. Molnár-Gábor, F. Artificial Intelligence in Healthcare: Doctors, Patients and Liabilities. in *Regulating Artificial Intelligence* (eds. Wischmeyer, T. & Rademacher, T.) 337–360 (Springer International Publishing, 2020). doi:10.1007/978-3-030-32361-5_15.
103. Reddy, S., Fox, J. & Purohit, M. P. Artificial intelligence-enabled healthcare delivery. *J. R. Soc. Med.* **112**, 22–28 (2019).
104. Blasimme, A. & Vayena, E. The Ethics of AI in Biomedical Research, Patient Care, and Public Health. in *The Oxford Handbook of Ethics of AI* (eds. Dubber, M. D., Pasquale, F. & Das, S.) 702–718 (Oxford University Press, 2020). doi:10.1093/oxfordhb/9780190067397.013.45.
105. Bedoya, A. D. *et al.* A framework for the oversight and local deployment of safe and high-quality prediction models. *J. Am. Med. Inform. Assoc.* **29**, 1631–1636 (2022).
106. Bozkurt, S. *et al.* Reporting of demographic data and representativeness in machine learning models using electronic health records. *J. Am. Med. Inform. Assoc.* **27**, 1878–1884 (2020).
107. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A. & Shah, N. H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Inform. Assoc.* **27**, 2011–2015 (2020).
108. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
109. Morley, J. *et al.* The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* **260**, 113172 (2020).
110. Blasimme, A. & Vayena, E. Becoming partners, retaining autonomy: ethical considerations on the development of precision medicine. *BMC Med. Ethics* **17**, 67 (2016).
111. Grote, T. & Berens, P. On the ethics of algorithmic decision-making in healthcare. *J. Med. Ethics* **46**, 205–211 (2020).
112. Hofmann, B. & Stanak, M. Nudging in screening: Literature review and ethical guidance. *Patient Educ. Couns.* **101**, 1561–1569 (2018).
113. Andorno, R. The right not to know: an autonomy based approach. *J. Med. Ethics* **30**, 435–439 (2004).

114. Schwartz, P. H. & Meslin, E. M. The Ethics of Information: Absolute Risk Reduction and Patient Understanding of Screening. *J. Gen. Intern. Med.* **23**, 867–870 (2008).

115. Green, S. & Vogt, H. Personalizing Medicine: Disease Prevention in silico and in socio. *HUMANA MENTE J. Philos. Stud.* **9**, 105–145 (2016).

116. Morley, J. & Floridi, L. The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. *Sci. Eng. Ethics* **26**, 1159–1183 (2020).

117. Kerasidou, A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bull. World Health Organ.* **98**, 245–250 (2020).

118. Chin-Yee, B. & Upshur, R. Three Problems with Big Data and Artificial Intelligence in Medicine. *Perspect. Biol. Med.* **62**, 237–256 (2019).

119. McDougall, R. J. Computer knows best? The need for value-flexibility in medical AI. *J. Med. Ethics* **45**, 156–160 (2019).

120. Heyen, N. B. & Salloch, S. The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. *BMC Med. Ethics* **22**, 112 (2021).

121. Bartoletti, I. AI in Healthcare: Ethical and Privacy Challenges. in *Artificial Intelligence in Medicine* (eds. Riaño, D., Wilk, S. & ten Teije, A.) vol. 11526 7–10 (Springer International Publishing, 2019).

122. Fritzsche, M.-C. *et al.* Ethical layering in AI-driven polygenic risk scores—New complexities, new challenges. *Front. Genet.* **14**, 1098439 (2023).

123. Vogt, H., Green, S., Ekstrøm, C. T. & Brodersen, J. How precision medicine and screening with big data could increase overdiagnosis. *BMJ* l5270 (2019) doi:10.1136/bmj.l5270.

124. McCartney, M. *et al.* Why 'case finding' is bad science. *J. R. Soc. Med.* **113**, 54–58 (2020).

125. Rubeis, G. Liquid Health. Medicine in the age of surveillance capitalism. *Soc. Sci. Med.* **322**, 115810 (2023).

126. Verheij, R. A., Curcin, V., Delaney, B. C. & McGilchrist, M. M. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J. Med. Internet Res.* **20**, e185 (2018).

127. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern. Med.* **178**, 1544 (2018).

128. Cirillo, D. *et al.* Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digit. Med.* **3**, 81 (2020).

129. Abettan, C. Between hype and hope: What is really at stake with personalized medicine? *Med. Health Care Philos.* **19**, 423–430 (2016).

130. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **27**, 2020–2023 (2020).

131. Gray, M., Lagerberg, T. & Dombrádi, V. Equity and Value in 'Precision Medicine'. *New Bioeth.* **23**, 87–94 (2017).

132. McCradden, M. D. *et al.* Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *J. Am. Med. Inform. Assoc.* **27**, 2024–2027 (2020).

133. Parikh, R. B., Obermeyer, Z. & Navathe, A. S. Regulation of predictive analytics in medicine. *Science* **363**, 810–812 (2019).

134. Paulus, J. K. & Kent, D. M. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Npj Digit. Med.* **3**, 99 (2020).

135. the Precise4Q consortium *et al.* Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 310 (2020).

136. Price, W. N. Big data and black-box medical algorithms. *Sci. Transl. Med.* **10**, eaao5333 (2018).