bmjmedicine

How to read a paper involving artificial intelligence (AI)

Paul Dijkstra [1], Trisha Greenhalgh [2], Yosra Magdi Mekki,[3] Jessica Morley[4]

[1]Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[3]Qatar University College of Medicine, Doha, Qatar
[4]Yale University, New Haven, Connecticut, USA
Correspondence to: Professor Trisha Greenhalgh, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK; trish.greenhalgh@phc.ox.ac.uk

Cite this as: *BMJMED* 2025;4. doi:10.1136/bmjmed-2025-001394

Received: 4 February 2025
Accepted: 19 March 2025

ABSTRACT

This paper guides readers through the critical appraisal of a paper that includes the use of artificial intelligence (AI) in clinical settings for healthcare delivery. A brief introduction to the different types of AI used in healthcare is given, along with some ethical principles to guide the introduction of AI systems into healthcare. Existing publication guidelines for AI studies are highlighted. Ten preliminary questions to ask about a paper describing an AI based decision support algorithm are suggested.

Introduction

Increasingly, clinical research papers describe the use of artificial intelligence (AI), and much has been written about the potential of AI to revolutionise healthcare.[1] Many examples exist of AI being rolled out into healthcare delivery (examples in boxes 1 and 2, and recent systematic reviews[2 3]), from the use of a machine learning algorithm to distinguish between a cough caused by a pulmonary tuberculosis[4] and non-tuberculosis respiratory condition (box 1) to using generative AI to improve decision making at the clinical frontline (box 2).[1 3] AI, however, is not a panacea, and it raises both practical and ethical challenges. AI applications should undergo rigorous research, and those reading research on AI need to be able to evaluate it to understand the effect of AI applications on patients and healthcare systems.

In this paper, we aim to provide novice readers with an introduction to the many uses of AI in healthcare and how to begin to critically appraise these studies. Our objectives were to: introduce and define key concepts; outline some ethical principles for health related use of AI; give examples of reporting guidelines and checklists for appraising papers reporting AI studies; and propose some initial questions to ask about papers describing machine learning decision support systems (one of the best established uses of AI in clinical practice). This approach should be used to complement, rather than replace, the principles of critical appraisal for papers that do not include the study of AI in clinical care.[5]

What is artificial intelligence?

AI is a transformative technology capable of completing tasks that typically require human intelligence. AI is also an interdisciplinary area of research that spans, among other disciplines, computer science, psychology, linguistics, and philosophy. AI in healthcare is divided into two broad categories:

- Artificial narrow intelligence refers to machine learning algorithms that can recognise patterns in large datasets. Artificial narrow intelligence is useful for solving text, voice, or image based classification and clustering problems, excelling at a precisely defined, single task (eg, playing chess).
- Artificial general intelligence refers to AI applications that can reason, argue, memorise, and solve problems. Artificial general intelligence is sometimes referred to as human level AI, because it is considered to display a cognitive capacity which approximates to that of a human being.

Whereas artificial narrow intelligence is already having a major effect on medical practice, including diagnosis, risk prediction and prognosis, image interpretation, surgical practice (eg, computer vision for laparoscopic cholecystectomy), and drug discovery, few robust examples exist of artificial general intelligence delivering benefits in clinical practice.[6] Also important in healthcare, but beyond the scope of this paper, is the many administrative functions of AI (eg, speech recognition products that allow the user to dictate a letter and see it typed as you speak).

From a computational perspective, the diverse medical and healthcare uses of AI can be classified into five broad categories:

- prediction (using historical data to predict the likelihood of future events);
- classification (eg, of images into normal or abnormal);
- association (finding underlying relations between variables for the purpose of enhancing prediction);
- regression (assessing the strength of a relation between one variable and a series of other potentially related variables); and
- optimisation (mostly, administrative tasks).

The glossary (box 3) lists some key terminology and definitions, and we expand on some of these below.

KEY MESSAGES

⇒ Papers describing research on or with artificial intelligence (AI) are now commonplace
⇒ Some forms of AI have an established place in clinical practice whereas others are speculative
⇒ A preliminary framework and set of questions for appraising a paper describing an AI based decision support algorithm are described

BOX 1 | EXAMPLE OF USE OF ARTIFICIAL INTELLIGENCE TO DISTINGUISH BETWEEN TUBERCULOSIS AND NON-TUBERCULOSIS COUGH (MACHINE LEARNING ALGORITHM)

A team of researchers collected data from 149 individuals with pulmonary tuberculosis and 46 with non-tuberculosis respiratory conditions in Nairobi, Kenya.[4] They recorded >33 000 passive coughs and 1600 forced coughs in a controlled environment, designed to minimise background noise and environmental variability. This set-up ensured that image classification (a key aspect of computer vision) would focus on cough specific features rather than external factors. The system analysed scalograms (ie, visual representations of the frequency content in cough sounds). The model was tasked with learning which features were uniquely associated with tuberculosis. The researchers developed a cough classifier built on ResNet18 architecture, a deep learning model commonly used in computer vision tasks, such as image classification. The model analysed scalograms to identify patterns that might differentiate coughs related to tuberculosis from other coughs. These results indicate that AI can identify diagnostic features in cough sounds associated with tuberculosis, especially in patients with greater severity of disease. The system had effectively learnt to "listen" to a cough in ways that even trained clinicians could not.

Big data

Big data refers to large or complex datasets that would be impossible to analyse without advanced computer power.[3] Because the size of cohort studies and the number of variables to be analysed has increased, calculations can no longer be done on a desktop spreadsheet. Cloud computing and the rapid growth in computer processing power are the key drivers of the AI revolution.

One example of a big data platform is Cosmos, developed by Epic in the US.[7] Cosmos is the largest available database of electronic health record data, integrating the inpatient and outpatient records of 289 million patients, providing data to researchers on more than 14.8 billion clinical encounters (including 7.2 billion face-to-face visits) across 37 700 clinics and 1626 hospitals. The resultant dataset has a wide range of data, including personal characteristics, vital signs, drug treatments, laboratory results, procedures, diagnoses, encounters, patient generated data, and clinical domain specific data, as well as data on the social determinants of health.

BOX 2 | EXAMPLE OF USE OF GENERATIVE ARTIFICIAL INTELLIGENCE IN CLINICAL PRACTICE

In *The AI Revolution in Medicine*,[1] Isaac Kohane describes how he was called to consult on a newborn baby with ambiguous genitalia. Was this a chromosomal male with underdeveloped genitalia or a chromosomal female who had been androgenised in utero? What should he tell the parents? A fast decision was needed for both medical and social reasons. Kohane put this clinical query into the newly developed GPT-4, an artificial intelligence (AI) digital assistant, accessible from a smartphone app. Within seconds, he got a response listing several possible diagnoses: congenital adrenal hyperplasia, androgen insensitivity syndrome, gonadal dysgenesis, and pituitary or hypothalamic dysfunction. For each option, the digital assistant had suggested some further tests to help exclude or recognise the condition and identify subtypes. The suggestions allowed Kohane to do a focused search of the literature, perform a panel of specialised tests on the infant, and establish a firm (and treatable) diagnosis in only a few days. This anecdotal example shows the potential of AI to complement (although not replace) clinical assessment.

Machine learning: many current benefits

Machine learning is a subtype of AI; models or algorithms learn patterns from data, rather than being programmed with rules.[8] Various types (some of which overlap) include supervised learning, unsupervised learning, reinforcement learning, deep learning, and zero shot learning (box 3 has definitions and box 1 has an example). A comprehensive overview of machine learning applications in healthcare is beyond the scope of this paper, but problems such as data leakage, undercutting, and data poisoning could result in clinically significant errors in machine learning applications. Leakage, for example, leads to inflated model performance and decreased reproducibility.[9] A classic example is the inclusion of the patient's identification number as a predictor. Machine learning algorithms will learn, for example, that similar oncology hospital identification numbers have a higher probability of cancer.[10] Furthermore, as with any clinical research, if a machine learning study is undertaken on a biased sample (eg, one that is less sick or less complex than the wider population of patients with the condition), the findings will be untrustworthy and might lead to harm.

Generative AI and its potential for the future

Generative AI, discussed in more detail elsewhere,[11] includes large language models (LLMs)[12] and large multimodal models (LMMs)[12 13] (box 3). LLM AI systems use algorithms trained on billions of words from articles, books, and other text based internet content to generate language ("talk") like humans. LMMs do the same but accept many types of data input, such as text, images, audio, and video, and sometimes other data types (eg, sensory data).[12 13] LMMs generate diverse outputs that are not necessarily related to the type of data fed to the algorithm (eg, images fed to the algorithm could generate text, or vice versa).[14] LLMs and LMMs are increasingly used in medicine to retrieve knowledge, support clinical decision making, summarise key findings, and triage patients, for example. A recent paper summarised the extent to which LLMs "encode" clinical knowledge, including their limitations and suggestions for future research.[15] An inherent problem with LLMs and LMMs is their tendency to "hallucinate" (eg, cite papers that do not exist) and recommend courses of action that would not be in the patient's best interests.[1] Medical LLMs are also vulnerable to data poisoning, when corrupted with medical misinformation, resulting in medically harmful text.[16]

BOX 3 | GLOSSARY: TERMINOLOGY AND DEFINITIONS

Artificial intelligence (AI)

Various definitions exist. One (from IBM) is "technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity, and autonomy."[35]

AI system

A system that incorporates AI (eg, consists of the AI algorithm and its supporting software and hardware platforms).

Big data

Usually defined as data characterised by very high volume, velocity, and variety requiring special technology and analytical methods to transform it into value. Big data has eight key properties: great variety, high velocity, challenge on veracity, challenge on all aspects of the workflow, challenge on computational methods, challenge on extracting meaningful information, challenge on sharing data, and challenge on finding human experts.[36] Some definitions of big data use only volume (eg, if $\log(n \times p)$ is ≥ 7).[36]

Calibration (and miscalibration)

Calibration refers to how well the predicted probabilities of a model align with the actual likelihood of events. Miscalibration refers to situations where this alignment is poor. One example of miscalibration is undercutting (ie, when a model's predictions are consistently lower than true values). This might occur if the training set is biased, the model used is too simple, or features used to train the model are not representative of the underlying phenomenon.

ChatGPT

Chat Generative Pre-trained Transformer, a commercial digital assistant (accessible from a smartphone app, for example). ChatGPT is an inherently dialogic text or image based interface to a GPT architecture large language model. A free basic version is available as well as a more sophisticated version for a monthly fee.

Cloud computing

The on-demand availability of computing resources (such as storage and infrastructure) over the internet.

Computer processing power

The ability of a computer to process information or the speed at which it can process information.

Conversational AI

An AI system which replicates human intelligence in conversation with a human. Examples include the dialogic form used in ChatGPT.

Data leakage

Occurs when some of the data used for training is leaked into the testing or calibration datasets. This leakage leads to high performance on the training set (and sometimes also the validation data), but the model will perform much less well in production.

Deep learning

A type of machine learning that uses a layered structure of artificial neural networks. These networks consist of input, hidden, and output layers, with information processed through methods such as forward and backward propagation to adjust weights and optimise performance. The topology of the networks includes convolutional, recurrent, and transformer architectures. An example is AlphaFold2 (AF2), an AI system that can predict the three dimensional structures of proteins from amino acid sequences. This system has revolutionised drug discovery and won its architects the Nobel Prize for chemistry in 2024.[37 38]

Data poisoning

A malicious attack in which an adversary intentionally introduces harmful or misleading data into a training dataset.

Fairness

Fairness in AI aims to ensure that AI systems treat individuals and groups equitably, without unjustified discrimination or bias. Fairness checks are processes and methods used to identify and mitigate biases in AI models.

Generative AI

A subset of AI technologies that generate new content such as text, images, music, speech, video, or code by learning the patterns and structure of large amounts of training data.

Intended AI use

The use for which an AI system is intended (eg, targeted medical condition, patient and user populations, and use environment).

Large language models (LLMs)

AI models that use computational AI algorithms with text data as input to generate language that resembles that produced by humans.

Large multimodal models (LMMs)

AI models that have the ability to accept one or more types of data inputs (eg, text, videos, and images) and generate diverse outputs that are not limited to the type of data inputted.

Machine learning

A area of computer science where models and algorithms learn patterns from data, rather than being programmed with rules. For example, systems such as CHICA (Child Health Improvement through Computer Automation) use machine learning principles and expert system methodologies to improve paediatric care.

(Continued)

BOX 3 | (CONTINUED)

Reinforcement learning

A type of machine learning where an agent learns to interact with an environment by trial and error, by optimising its decisions through trial and error, guided by the optimisation of cost or reward functions. This technique is powerful in training AI agents to make decisions in complex environments by learning from experience. Examples include optimising medical treatment plans (eg, personalising insulin dosing for patients with diabetes) and developing adaptive therapeutic strategies in healthcare.

Supervised learning

A type of machine learning where the algorithm is trained on a dataset with labelled examples, with methods such as linear regression (eg, predicting blood pressure based on age, weight, and drug treatment dosage) or logistic regression (eg, determining the likelihood of a patient having diabetes based on test results and a family history). These labels provide the correct output or target variable for each input. Examples include disease risk prediction, diagnostic imaging classification, and personalised treatment recommendations.

Unsupervised learning

A type of machine learning where the algorithm is trained on a dataset without any labels. The algorithm discovers patterns, structures, or relationships in the data, with methods such as clustering (eg, grouping patients by similar disease characteristics) or dimensionality reduction (eg, reducing the complexity of gene expression data for analysis). Example applications include disease clustering (eg, identifying subtypes of breast cancer), early detection of medical conditions, and drug discovery.

Zero shot learning

The ability of an algorithm to perform tasks or make predictions on data that it has not encountered during training, without requiring additional fine tuning.

One LLM, ChatGPT (Chat Generative Pre-trained Transformer), has passed the US Medical Licensing Examination[17] and has been used to answer common patient questions (eg, about colonoscopy[18] in research studies). Because they operate in a conversational format, LLMs and LMMs can feel like interacting with a human or human-like agent (ie, you ask a question, the LLM replies, you explain why the response is not quite what you wanted, and so on). Also, because these technologies are designed to find and take account of context rather than focusing only on the inputted text, the models can potentially adapt what they say (and how they say it) to accommodate different patient personalities and levels of health literacy. These models also have the ability to respond in the patient's preferred language. Thus generative AI could potentially overcome some of the brittleness of previous generation digital health applications and (somewhat paradoxically) help humanise the interface between the patient and the system. Although about 20% of general practices in the UK are already experimenting with tools such as ChatGPT,[19] at the time of writing, the potential rather than the actual benefits for patients are being discussed.

Ethical challenges of AI applications in health

What comes out of an AI system depends on how it is built, reminiscent of the common adage "rubbish in, rubbish out." AI systems trained on poor quality data are likely to be biased.[20] Examples of poor quality data are data that are negatively affected by problems such as missingness (ie, when key personal or clinical variables are unrecorded), non-representativeness (ie, when key groups of the population targeted by the AI system are completely absent or represented in too small numbers), or misclassification (ie, when data reflects incorrect assumptions about patients or other users). This approach means that AI systems might perform better (ie, more accurately identify or exclude a condition) in people who were more fully and more accurately reflected in the training dataset than those who were unrepresented or inaccurately reflected in the training dataset.[21] Over time, this biased performance can harm groups that are already under-represented and marginalised.[22]

Mitigating the consequences of bias is challenging because bias can also appear in other parts of the AI development pipeline, as well as in the training dataset, making it difficult to detect its cause. Biases can, for example, be introduced during the design of the model due to the selection or weighting of different variables or, after implementation, as a result of interaction with social bias (ie, unconscious or conscious discrimination by human clinicians), or because of dataset or population drift (ie, when the input data used in frontline care changes or the make-up of the target population changes). Hence continuing to monitor the performance and effect of AI systems after they have been deployed is important.

The ethical implications of AI reach beyond bias. Whatever its data processing ability, a machine is still a machine, with no semantic understanding, and aspects of medicine and healthcare (eg, compassion, comfort, and care) will, arguably, always require human input. As Hicks et al have argued in relation to generative AI, "the models are in an important way indifferent to the truth of their outputs".[23] This observation is one reason

why AI applications are, for the most part, designed to be used by a human who has professional training, and not substitute for that human.

Use of AI applications requires human qualities, such as humility and circumspection. In a recent review, Messeri and Crockett discussed four visions of AI: AI as oracle, AI as arbiter, AI as quant, and AI as surrogate.[24] The authors warned that these various roles are cognitive traps that can produce various illusions: the illusion of explanatory depth (assuming that the explanation produced by AI is more profound than it actually is); the illusion of exploratory breadth (assuming that the AI model has covered all possible hypotheses relevant to the question when it has actually covered only a limited number); and the illusion of objectivity (assuming that the AI model has produced an unbiased view from nowhere when in reality it reflects, and might even magnify, the various biases inherent in the published literature on a topic).

For all of these reasons, ethical considerations and human rights must be central to the design, development, and implementation of AI tools. The World Health Organization has endorsed six key ethical principles for the use of AI for health (box 4), and also reviewed the complex topic of governance of generative AI models.[25]

BOX 4 | WORLD HEALTH ORGANIZATION'S (WHO) SIX ETHICAL PRINCIPLES FOR USE OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE

Summarised from WHO guidance on ethics and governance of AI in healthcare[25]

1. Protecting human autonomy. Humans should remain in control of medical decisions, and people should understand how artificial intelligence (AI) is used in their care (including how their privacy and confidentiality are protected).
2. Promoting human wellbeing and safety, and the public interest. AI should not harm people. The designers of AI technology should comply with regulatory requirements for safety, accuracy, and efficacy.
3. Ensuring transparency, explainability, and intelligibility. AI technology should be understandable to developers, healthcare professionals, patients, users, and regulators "according to the capacity of those to whom they are explained".
4. Fostering responsibility and accountability. Patients and clinicians should evaluate the development and deployment of AI technologies. This approach should include mechanisms for questioning and redress for individuals and groups that are adversely affected by decisions based on algorithms.
5. Ensuring inclusiveness and equity. AI for health should be designed "to encourage the widest possible appropriate, equitable use and access, irrespective of age, sex, gender, income, race, ethnic group, sexual orientation, ability, or other characteristics protected under human rights codes". AI technologies should not encode biases to the disadvantage of identifiable groups (especially already minoritised groups; ie, fairness, box 3).
6. Promoting AI that is responsive and sustainable. All AI role players (designers, developers, and users) should "continuously, systematically, and transparently" assess AI applications during actual use. Two aspects are important for sustainable AI systems: firstly, their environmental consequences should be minimal; and secondly, their effect on the workplace, including workplace disruptions, training of healthcare workers, and potential job losses should be dealt with by governments and companies.

Ten questions to ask about an AI based decision support algorithm

Box 5 lists 10 questions to keep in mind when appraising a paper describing AI in decision support. Many of these questions have wider relevance to other types of AI tools. In preparing these 10 questions, we have used several AI quality tools and reporting guidelines.[14 26–30]

Even when a structured critical appraisal based on the questions in box 5 confirms that the AI study was done rigorously and its findings can be trusted (internal validity), a further question is, does the AI technology or intervention work in (clinical) practice (external validity)? Patients attending clinics might not be comparable with the patient sample on which the algorithm was trained, and the staff using the technology might differ in important ways from those in the study.

Reporting guidelines for AI papers: further detail

The literature already includes many reporting guidelines and checklists, but so far no validated quality appraisal tools. Publications for authors on what items to report in papers describing AI applications are now commonplace. Table 1 summarises selected examples of these guidelines. Numerous checklists have been developed to help the reader appraise such papers (these focus not only on what should be reported but on how to decide whether the researchers have dealt with each item to a sufficiently high standard). The APPRAISE-AI (APPRAISAL of AI studies for Clinical Decision Support) tool, for example, was developed to evaluate the methodological and reporting quality of 28 clinical AI studies,[31] although many other frameworks have been developed in recent years (a 2024 review identified 26, including at least nine reporting checklists).[32]

Reporting guidelines are a good starting point when evaluating a paper, but they should not be used uncritically as quality checklists. The correct reporting guideline for the study design of the paper being appraised should be selected, for example TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis)+AI for prediction model studies, STARD (Standards for Reporting of Diagnostic Accuracy Studies)-AI for diagnostic accuracy studies, and SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials)-AI or CONSORT (Consolidated Standards of Reporting Trials)-AI for randomised controlled trials. All of these guidelines are AI extensions to established reporting guidelines (eg, with additional items to look at potential sources of bias specific to AI systems).[33]

BOX 5 | TEN QUESTIONS TO ASK ABOUT A PAPER THAT REPORTS AN ARTIFICIAL INTELLIGENCE (AI) BASED DECISION SUPPORT SYSTEM

1. **What was the study design and (excluding the AI) aspects for now) does it meet established criteria for methodological rigour?**

As with all research, studies of AI systems should be conducted with systematic methods, on samples that are large enough and representative enough to produce trustworthy findings. For example, if a randomised controlled trial was appropriate to the research question, was a randomised controlled trial done (and if not, what were the limitations of the non-randomised design)? Did the sample of patient participants reflect the wider population with the disease? If an intervention and a control arm were included, were participants in both arms comparable at baseline in terms of, for example, age distribution, gender balance, and disease severity? Were assessors blinded to the allocation arms? AI studies typically have two key groups of participants: the patients on which the AI algorithm was trained and the users of the AI system (consider their baseline characteristics and how they were familiarised with the AI system). This scenario is similar to surgical innovation studies where both patient and operator characteristics are important and should be reported.[39]

2. **What was the intended clinical use of the AI system (ie, what decisions is it intended to support, by whom, and with what purpose)?**

Be clear about the purpose of the technology. The AI system might, for example, be targeted for use in patients with suspected breast cancer and used by radiologists, with the goal of increasing detection rates of breast cancer and reducing false positive screens.

3. **What type of computational task was the AI system designed to support?**

The computational task is likely to be one or more of the following: prediction (eg, estimating the risk of a person developing a specific health outcome), classification (eg, estimating the presence or absence of a disease or risk state), association (eg, drug discovery or new risk factors for a specific health condition), regression (estimating the likelihood of a patient having a condition, given particular risk factors), or optimisation (eg, improving the efficiency of administrative tasks).

4. **What AI system was used and how?**

Look for a description of the AI system, especially its version, the underlying algorithm, supporting hardware, and software (if relevant). Which data were used as inputs for the AI system? How did the research team acquire the data? What method was used to input the data into the AI system? What preprocessing was applied and how were missing or low quality data handled? Finally, how were the outputs of the AI system presented to the users? How information is displayed influences how users interact with an AI system.[40]

5. **Where was the AI system** located **in the clinical workflows or pathways?**

It matters when users see AI system information. Anchoring bias, a form of cognitive bias,[41] is possible when users see the AI system recommendations in concurrent reading mode (ie, at the same time as they receive other information). In second reading mode, users see the AI system recommendations after they have made the decision about a patient and re-evaluate their initial decision based on the new information.

6. **What was the approach to possible errors in the AI system? Did the authors report and discuss any safety concerns or instances of harm?**

Three main categories of errors and malfunctions should be considered (look for this information in the methods section of the paper) and reported in the results section (eg, rate, causes, and effect on patient care of AI system errors or malfunctions):

a. algorithm errors (eg, AI system did not detect all cancers in women's digital mammograms)[41];
b. malfunction of the supporting software or hardware (eg, the AI system failed to produce a recommendation because of a problem with data extraction); and
c. user errors (eg, the clinician inputted inaccurate patient details or applied the AI system to a medical indication it was not designed for).

Safety assessment is a continuous process that occurs before, during, and after a clinical study because new risks or harms might be uncovered after implementation of the AI system in clinical settings. The greatest risks after implementation of machine learning applications are related to dataset drift (or population drift), resulting in declining performance over time that can be difficult to detect.

7. **How did the authors deal with human factors?**

Human factors (also called ergonomics, or how the design features of technologies affect whether and how humans interact with it, including factors such as situation awareness, workload, and techno-stress) can make or break AI systems in healthcare.[42] The paper should describe and seek to understand and explain key human technology interactions.

8. **How transparent were the authors about the data and code used to train and validate their AI system?**

The authors should share a description of the data (eg, data sheets for datasets), annotated to explain what each element does, to allow readers to follow this section even when not trained in the technical aspects of AI. The code should be available (eg, in a supplementary file).

9. **How did the authors deal with the ethical use of the AI system?**

The authors should describe, for example,

a. what techniques they used to detect, quantify, and mitigate bias in the algorithmic outputs of the AI system (eg, through algorithmic fairness, with adjustments to correct for bias);
b. how they dealt with privacy and security, and whether their approach was appropriate and sufficient for the type of data used; and
c. what patient facing information and explanations were included and whether these were adequate for patients to make informed decisions.

10. **Did the research incorporate multiple types of expertise?**

(Continued)

BOX 5 | (CONTINUED)

Research into AI systems for clinical use needs more than technical experts. A diverse team of AI scientists, clinicians, and patient partners is the most appropriate to counteract the risk of "monocultures of knowing and knowers".[24] Clinicians and patients should be actively involved in building and testing any AI tool aimed at clinical care.[43] In question (2), for example, asking whether patients and clinicians were involved in determining if this task was appropriate to delegate to an algorithm is important.

See text for additional sources.

At the time of writing, no reporting standards exist for studies evaluating the performance of LLM linked chatbots when providing clinical advice. A group of researchers are currently developing CHART (chatbot assessment reporting tool) to ensure transparency when writing up research on LLM linked chatbots that summarise health evidence and provide clinical advice.[34] This assessment tool will be a key preliminary step to developing critical appraisal tools for evaluating such studies. Publication of CHART can be tracked in Google Scholar.

To show how to derive critical appraisal questions from existing reporting guidelines, we have adapted the DECIDE (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence)-AI checklist for evaluating AI based decision support systems (table 2). AI based decision support presents some methodological challenges, including: that these systems are complex interventions which are used (or not) by people as part of work routines and pathways in a wider health ecosystem involving human choices and judgments; that any AI application is continually changing through upgrades and system learning; and that features of the populations on which the algorithm is trained might introduce biases that could generate inequities.[14 31] Most of these challenges also apply to other types of AI study, for example, how humans and machines "collaborate" while protecting the important ethical principle of human autonomy, and how AI systems could harm (eg, algorithm bias and data breaches).

Table 1 | Selected reporting guidelines for papers that include artificial intelligence (from Vasey et al[14] and Kolbinger et al[32])

Name	Stage	Study design	Comments
TRIPOD+AI[28]	Preclinical development	Prediction model evaluation	Extension of TRIPOD and used to report development, validation, and updates of diagnostic and prognostic prediction models (diagnostic or prognostic)
STARD-AI[27] (protocol paper)	Preclinical development, offline validation	Diagnostic accuracy studies	Extension of STARD and used to report diagnostic accuracy studies
DECIDE-AI[14]	Early live clinical evaluation	Various (eg, prospective cohort studies, non-randomised controlled trials)	Standalone guideline that is used to report the early evaluation of AI systems as an intervention in live clinical settings (guideline used for all study designs and any AI system modality; eg, diagnostic, prognostic, or therapeutic). This guideline focuses on clinical use, safety, and human factors
SPIRIT-AI[26]	Comparative prospective evaluation	Randomised controlled trials (protocol)	Extension of SPIRIT. Used to report the protocols of randomised controlled trials evaluating AI systems as interventions
CONSORT-AI[29]	Comparative prospective evaluation	Randomised controlled trials	Extension of CONSORT and used to report large scale randomised controlled trials evaluating AI systems as interventions (for any AI system modality; eg, diagnostic, prognostic, or therapeutic). This guideline focuses on effectiveness and safety
CLAIM[30]	All AI in medical imaging studies	Studies reporting medical imaging	A checklist to guide reporting of AI in medical imaging
CHART[34]	All studies assessing the use of chatbots in healthcare	Chatbot assessment studies	CHART currently being developed
TRIPOD-LLM[44]	Studies that are developing, tuning, prompt engineering, or evaluating an LLM	LLM studies	Extension of the TRIPOD+AI statement. Looking at the unique challenges of LLMs in biomedical applications
CHEERS-AI[45]	Studies that report an economic evaluation of an intervention that uses AI to perform its function	Economic evaluation studies	CHEERS-AI checklist is intended to standardise reporting of economic evaluations of health technologies that use AI

AI, artificial intelligence; CHART, chatbot assessment reporting tool; CHEERS, Consolidated Health Economic Evaluation Reporting Standards; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CONSORT, Consolidated Standards of Reporting Trials; DECIDE, Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence; LLM, large language model; SPIRT, Standard Protocol Items: Recommendations for Interventional Trials; STARD, Standards for Reporting of Diagnostic Accuracy Studies; TRIPOD, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis.

Table 2 | Simplified DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence) checklist (summarised with permission from Vasey et al[14])

Theme	Recommendation
Title	Include AI in the title
Abstract	Provide a structured summary
Introduction:	
Intended use	a. Describe what condition or conditions the AI is being used for, the intended patient population, and current standard practice b. Describe who will use the AI system, how it will fit into the care pathway, and what the anticipated effect will be
Objectives	State the study objectives
Ethics and governance	Link to the study protocol and give details of ethics approval
Methods:	
Participants	a. Describe how patient participants and users of the AI system were recruited, including what inclusion and exclusion criteria were used (for patients, at both patient and data level), and justify the sample sizes b. Describe how users were taught to use the AI system
AI system	a. Describe the AI system (include version and type of underlying algorithm). Describe and reference the characteristics of the patient population on which the algorithm was trained (and its performance in preclinical development and validation studies) b. Identify and describe input data (eg, including how data were acquired and entered) c. Describe the AI system outputs and how they were presented to users (add image if possible)
Implementation	Describe the settings in which the AI system was evaluated, including details of clinical workflow, who made actual clinical decisions, and how
Outcomes	Specify primary and any secondary outcomes measured
Safety and errors	Describe how errors or malfunctions were defined and identified, and how risks to patient safety were identified, analysed, and minimised
Human factors	Describe the human factor tools, methods, or frameworks used, the used cases considered, and the users involved
Analysis	Describe and justify how data were analysed, including statistical tests
AI ethics	Give details of any specific methodologies used to ensure algorithmic fairness or other ethical goals
Patient involvement	Say how patients were involved in any aspect of the study
Results:	
Participants	Give details of both the patient population and users of the AI system
Implementation	Report on how much the AI system was used, whether users adhered to the protocol, and any changes to workflow or care pathway
Main findings	Report on the primary and secondary outcomes, with subgroup analyses if appropriate
Modifications	Report on any changes made to the AI platform or its hardware throughout the study
Human-computer agreement	Report on how closely the human user agreed with the AI system's recommendations, and explain and comment on any disagreements
Safety and errors	List and comment on major errors or malfunctions, including how common they were, whether they could be corrected, and effects on patient care. Report on any risks or harms to patients, including indirect harm
Human factors	Report on the usability evaluation (perhaps with a recognised framework)
Discussion:	
Conclusions about intended use	Discuss whether the study results support the intended use of the AI system in clinical settings
Safety and errors	Discuss the implications of the safety profile of the AI system, including the extent to which limitations might be mitigated
Strengths and limitations	Discuss the strengths and limitations of the study
Statements:	
Data availability	Say what data are available (including code) to others
Disclosures	Disclose any conflicts of interest (eg, commercial interests)

AI, artificial intelligence.

Conclusions

AI has huge potential in healthcare. AI, however, is not a panacea, and the risk of bias is widespread in AI studies. A structured approach based on the 10 questions to ask about an AI based decision support algorithm will help clinicians to distinguish robust and clinically important AI studies from those that add little clinical value, increase inequities, and harm instead of help.

Contributors TG and PD wrote the first draft of the paper. All authors contributed to refining the paper and all approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. TG is the guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests All authors have completed the ICMJE unifform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work; JM has received consulting fees from Owkin for general advice on ethical AI; JM is a member of the UK National Data Guardian panel and NHS England AI Advisory Board; PD is associate editor of the British Journal of Sports Medicine; TG is a member of Independent SAGE; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs

Paul Dijkstra http://orcid.org/0000-0003-3166-1357
Trisha Greenhalgh http://orcid.org/0000-0003-2369-8088

REFERENCES

1 Lee P, Goldberg C, Kohane I. The AI Revolution in Medicine: GPT-4 and Beyond. 1st edn. Hoboken: Pearson, 2023.
2 Younis HA, Eisa TAE, Nasser M, *et al*. A Systematic Review and Meta-Analysis of Artificial Intelligence Tools in Medicine and Healthcare: Applications, Considerations, Limitations, Motivation and Challenges. Diagnostics (Basel) 2024;14:109. 10.3390/diagnostics14010109
3 Abbaoui W, Retal S, El Bhiri B, *et al*. Towards revolutionizing precision healthcare: A systematic literature review of artificial intelligence methods in precision medicine. Informatics in Medicine Unlocked 2024;46:101475. 10.1016/j.imu.2024.101475
4 Sharma M, Nduba V, Njagi LN, *et al*. TBscreen: A passive cough classifier for tuberculosis screening with a controlled dataset. Sci Adv 2024;10:eadi0282. 10.1126/sciadv.adi0282
5 Greenhalgh T. How To Read A Paper - series homepage on bmj.com, Available: https://www.bmj.com/about-bmj/resources-readers/publications/how-read-paper
6 Rodman A, Beam AL, Manrai AK. Exploring the past—and future—of medical diagnosis and artificial intelligence. *NEJM AI* 2024;1:9:AIp2400707.
7 Anonymous. Epic cosmos. Available: https://cosmos.epic.com [Accessed 4 Apr 2025].
8 Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126. 10.1038/s41746-020-00333-z
9 Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y) 2023;4:100804. 10.1016/j.patter.2023.100804
10 Chiavegatto Filho A, Batista A, Dos Santos HG. Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on 'Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning'. J Med Internet Res 2021;23:e10969. 10.2196/10969
11 Yu P, Xu H, Hu X, *et al*. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. Healthcare (Basel) 2023;11:2776. 10.3390/healthcare11202776
12 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. Nat Med 2023;29:1930–40. 10.1038/s41591-023-02448-8
13 Clusmann J, Kolbinger FR, Muti HS, *et al*. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3:141. 10.1038/s43856-023-00370-1
14 Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022;377:e070904. 10.1136/bmj-2022-070904
15 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. Nature New Biol 2023;620:172–80. 10.1038/s41586-023-06291-2
16 Alber DA, Yang Z, Alyakin A, *et al*. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025;31:618–26. 10.1038/s41591-024-03445-1
17 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. 10.1371/journal.pdig.0000198
18 Lee T-C, Staller K, Botoman V, *et al*. ChatGPT Answers Common Patient Questions About Colonoscopy. Gastroenterology 2023;165:509–11. 10.1053/j.gastro.2023.04.033
19 Blease CR, Locher C, Gaab J, *et al*. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform 2024;31:e101102. 10.1136/bmjhci-2024-101102
20 Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA 2019;322:2377–8. 10.1001/jama.2019.18058
21 Walsh CG, Chaudhry B, Dua P, *et al*. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open 2020;3:9–15. 10.1093/jamiaopen/ooz054
22 Aquino YSJ, Carter SM, Houssami N, *et al*. Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: a qualitative study of multidisciplinary expert perspectives. J Med Ethics 2023. 10.1136/jme-2022-108850
23 Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. Ethics Inf Technol 2024;26:38. 10.1007/s10676-024-09775-5
24 Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. Nature New Biol 2024;627:49–58. 10.1038/s41586-024-07146-0
25 World Health Organisation. Ethics and governance of artificial intelligence for health. Geneva WHO. Available: https://www.who.int/publications/i/item/9789240029200 [accessed 4 Apr 2025]
26 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2020;2:e549–60. 10.1016/S2589-7500(20)30219-3
27 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open 2021;11:e047709. 10.1136/bmjopen-2020-047709
28 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. 10.1136/bmj-2023-078378
29 Liu X, Rivera SC, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164. 10.1136/bmj.m3164
30 Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2020;2:e200029. 10.1148/ryai.2020200029
31 Kwong JCC, Khondker A, Lajkosz K, *et al*. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Netw Open 2023;6:e2335377. 10.1001/jamanetworkopen.2023.35377
32 Kolbinger FR, Veldhuizen GP, Zhu J, *et al*. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med (Lond) 2024;4:71. 10.1038/s43856-024-00492-0
33 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689. 10.1136/bmj.m689
34 Huo B, Cacciamani GE, Collins GS, *et al*. Reporting standards for the use of large language model-linked chatbots for health advice. Nat Med 2023;29:2988. 10.1038/s41591-023-02656-2
35 Anonymous. What is artificial intelligence (AI)?: IBM. 2025. Available: https://www.ibm.com/think/topics/artificial-intelligence [Accessed 31 Jan 2025].
36 Baro E, Degoul S, Beuscart R, *et al*. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int 2015;2015:639021. 10.1155/2015/639021
37 Varadi M, Anyango S, Deshpande M, *et al*. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50:D439–44. 10.1093/nar/gkab1061
38 Yang Z, Zeng X, Zhao Y, *et al*. AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther 2023;8:115. 10.1038/s41392-023-01381-z
39 Hirst A, Philippou Y, Blazeby J, *et al*. No Surgical Innovation Without Evaluation: Evolution and Further Development of the IDEAL

Framework and Recommendations. Ann Surg 2019;269:211–20. 10.1097/SLA.0000000000002794
40 Dudley JJ, Kristensson PO. A Review of User Interface Design for Interactive Machine Learning. ACM Trans Interact Intell Syst 2018;8:1–37. 10.1145/3185517
41 Freeman K, Geppert J, Stinton C, *et al*. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021;374:n1872. 10.1136/bmj.n1872
42 Sujan M, Pool R, Salmon P. Eight human factors and ergonomics principles for healthcare artificial intelligence. BMJ Health Care Inform 2022;29:e100516. 10.1136/bmjhci-2021-100516
43 Mekki YM. Physicians should build their own machine-learning models. Patterns (N Y) 2024;5:100948. 10.1016/j.patter.2024.100948
44 Gallifant J, Afshar M, Ameen S, *et al*. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025;31:60–9. 10.1038/s41591-024-03425-5
45 Elvidge J, Hawksworth C, Avşar TS, *et al*. Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI). Value Health 2024;27:1196–205. 10.1016/j.jval.2024.05.006