Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies

Myura Nagendran,[1] Yang Chen,[2] Christopher A Lovejoy,[3] Anthony C Gordon,[1,4] Matthieu Komorowski,[5] Hugh Harvey,[6] Eric J Topol,[7] John P A Ioannidis,[8] Gary S Collins,[9,10] Mahiben Maruthappu[3]

For numbered affiliations see end of the article.

Correspondence to:
M Nagendran, Intensive Care, St Mary's Campus, Imperial College London, Praed Street, London W2 1NY, UK myura.nagendran@imperial.ac.uk (or @MyuraNagendran on Twitter: ORCID 0000-0002-4656-5096)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;368:m689
http://dx.doi.org/10.1136/bmj.m689

Accepted: 11 February 2020

ABSTRACT

OBJECTIVE

To systematically examine the design, reporting standards, risk of bias, and claims of studies comparing the performance of diagnostic deep learning algorithms for medical imaging with that of expert clinicians.

DESIGN

Systematic review.

DATA SOURCES

Medline, Embase, Cochrane Central Register of Controlled Trials, and the World Health Organization trial registry from 2010 to June 2019.

ELIGIBILITY CRITERIA FOR SELECTING STUDIES

Randomised trial registrations and non-randomised studies comparing the performance of a deep learning algorithm in medical imaging with a contemporary group of one or more expert clinicians. Medical imaging has seen a growing interest in deep learning research. The main distinguishing feature of convolutional neural networks (CNNs) in deep learning is that when CNNs are fed with raw data, they develop their own representations needed for pattern recognition. The algorithm learns for itself the features of an image that are important for classification rather than being told by humans which features to use. The selected studies aimed to use medical imaging for predicting absolute risk of existing disease or classification into diagnostic groups (eg, disease or non-disease). For example, raw chest radiographs tagged with a label such as pneumothorax or no pneumothorax and the CNN learning which pixel patterns suggest pneumothorax.

REVIEW METHODS

Adherence to reporting standards was assessed by using CONSORT (consolidated standards of reporting trials) for randomised studies and TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) for non-randomised studies. Risk of bias was assessed by using the Cochrane risk of bias tool for randomised studies and PROBAST (prediction model risk of bias assessment tool) for non-randomised studies.

RESULTS

Only 10 records were found for deep learning randomised clinical trials, two of which have been published (with low risk of bias, except for lack of blinding, and high adherence to reporting standards) and eight are ongoing. Of 81 non-randomised clinical trials identified, only nine were prospective and just six were tested in a real world clinical setting. The median number of experts in the comparator group was only four (interquartile range 2-9). Full access to all datasets and code was severely limited (unavailable in 95% and 93% of studies, respectively). The overall risk of bias was high in 58 of 81 studies and adherence to reporting standards was suboptimal (<50% adherence for 12 of 29 TRIPOD items). 61 of 81 studies stated in their abstract that performance of artificial intelligence was at least comparable to (or better than) that of clinicians. Only 31 of 81 studies (38%) stated that further prospective studies or trials were required.

CONCLUSIONS

Few prospective deep learning studies and randomised trials exist in medical imaging. Most non-randomised trials are not prospective, are at high risk of bias, and deviate from existing reporting standards. Data and code availability are lacking in most studies, and human comparator groups are often small. Future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions.

STUDY REGISTRATION

PROSPERO CRD42019123605.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The volume of published research on deep learning, a branch of artificial intelligence (AI), is rapidly growing

Media headlines that claim superior performance to doctors have fuelled hype among the public and press for accelerated implementation

WHAT THIS STUDY ADDS

Few prospective deep learning studies and randomised trials exist in medical imaging

Most non-randomised trials are not prospective, are at high risk of bias, and deviate from existing reporting standards

Data and code availability are lacking in most studies, and human comparator groups are often small

Future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions

Introduction

The digitisation of society means we are amassing data at an unprecedented rate. Healthcare is no exception, with IBM estimating approximately one million gigabytes accruing over an average person's lifetime and the overall volume of global healthcare data doubling every few years.[1] To make sense of these big data, clinicians are increasingly collaborating with computer scientists and other allied disciplines to

make use of artificial intelligence (AI) techniques that can help detect signal from noise.[2] A recent forecast has placed the value of the healthcare AI market as growing from $2bn (£1.5bn; €1.8bn) in 2018 to $36bn by 2025, with a 50% compound annual growth rate.[3]

Deep learning is a subset of AI which is formally defined as "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction."[4] In practice, the main distinguishing feature between convolutional neural networks (CNNs) in deep learning and traditional machine learning is that when CNNs are fed with raw data, they develop their own representations needed for pattern recognition; they do not require domain expertise to structure the data and design feature extractors.[5] In plain language, the algorithm learns for itself the features of an image that are important for classification rather than being told by humans which features to use. A typical example would be feeding in raw chest radiographs tagged with a label such as either pneumothorax or no pneumothorax and the CNN learning which pixel patterns suggest pneumothorax. Fields such as medical imaging have seen a growing interest in deep learning research, with more and more studies being published.[6] Some media headlines that claim superior performance to doctors have fuelled hype among the public and press for accelerated implementation. Examples include: "Google says its AI can spot lung cancer a year before doctors" and "AI is better at diagnosing skin cancer than your doctor, study finds."[7 8]

The methods and risk of bias of studies behind such headlines have not been examined in detail. The danger is that public and commercial appetite for healthcare AI outpaces the development of a rigorous evidence base to support this comparatively young field. Ideally, the path to implementation would involve two key steps. Firstly, well conducted and well reported development and validation studies that describe an algorithm and its properties in detail, including predictive accuracy in the target setting. Secondly, well conducted and transparently reported randomised clinical trials that evaluate usefulness in the real world. Both steps are important to ensure clinical practice is determined based on the best evidence standards.[9-12]

Our systematic review seeks to give a contemporary overview of the current standards of deep learning research for clinical applications. Specifically, we sought to describe the study characteristics, and evaluate the methods and quality of reporting and transparency of deep learning studies that compare diagnostic algorithm performance with human clinicians. We aim to suggest how we can move forward in a way that encourages innovation while avoiding hype, diminishing research waste, and protecting patients.

Methods

The protocol for this study was registered in the online PROSPERO database (CRD42019123605) before search execution. The supplementary appendix gives details of any deviations from the protocol. This manuscript has been prepared according to the PRISMA (preferred reporting items for systematic reviews and meta-analyses) guidelines and a checklist is available in the supplementary appendix.[13]

Study identification and inclusion criteria

We performed a comprehensive search by using free text terms for various forms of the keywords "deep learning" and "clinician" to identify eligible studies. Appendix 1 presents the exact search strategy. Several electronic databases were searched from 2010 to June 2019: Medline, Embase, Cochrane Central Register of Controlled Trials (CENTRAL), and the World Health Organization International Clinical Trials Registry Platform (WHO-ICTRP) search portal. Additional articles were retrieved by manually scrutinising the reference lists of relevant publications.

We selected publications for review if they satisfied several inclusion criteria: a peer reviewed scientific report of original research; English language; assessed a deep learning algorithm applied to a clinical problem in medical imaging; compared algorithm performance with a contemporary human group not involved in establishing the ground truth (the true target disease status verified by best clinical practice); and at least one human in the group was considered an expert. We included studies when the aim was to use medical imaging for predicting absolute risk of existing disease or classification into diagnostic groups (eg, disease or non-disease). Exclusion criteria included informal publication types (such as commentaries, letters to the editor, editorials, meeting abstracts). Deep learning for the purpose of medical imaging was defined as computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction (in practice through a CNN; see box 1).[4] A clinical problem was defined as a situation in which a patient would usually see a medical professional to improve or manage their health (this did not include segmentation tasks, eg, delineating the borders of a tumour to calculate tumour volume). An expert was defined as an appropriately board certified specialist, attending physician, or equivalent. A real world clinical environment was defined as a situation in which the algorithm was embedded into an active clinical pathway. For example, instead of an algorithm being fed thousands of chest radiographs from a database, in a real world implementation it would exist within the reporting software used by radiologists and be acting or supporting the radiologists in real time.

Study selection and extraction of data

After removal of clearly irrelevant records, four people (MN, YC, CAL, Dina Radenkovic) independently screened abstracts for potentially eligible studies so that each record was reviewed by at least two people. Full text reports were then assessed for eligibility with disagreements resolved by consensus. At least two people (MN, YC, CAL) extracted data from study reports independently and in duplicate for each eligible study,

Box 1: Deep learning in imaging with examples

Deep learning is a subset of artificial intelligence that is formally defined as "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction."[4] A deep learning algorithm consists of a structure referred to as a deep neural network of which a convolutional neural network (CNN) is one particular type frequently used in imaging. CNNs are structurally inspired by the hierarchical arrangement of neurons within the brain. They can take many nuanced forms but the basic structure consists of an input layer, multiple hidden layers, and a final output layer. Each hidden layer responds to a different aspect of the raw input. In the case of imaging, this could be an edge, colour, or specific pattern.

The key difference between deep learning and other types of machine learning is that CNNs develop their own representations needed for pattern recognition rather than requiring human input to structure the data and design feature extractors. In plain language, the algorithm learns for itself the features of an image that are important for classification. Therefore, the algorithm has the freedom to discover classification features that might not have been apparent to humans (particularly when datasets are large) and thereby improve the performance of image classification.

CNNs use raw image data that have been labelled by humans in a process known as supervised learning. Each image is fed into the input layer of the algorithm as raw pixels and then processed sequentially through the layers of the CNN. The final output is a classification likelihood of the image belonging to a prespecified group.

Some examples from this review include the following:

- Feeding in raw chest radiographs tagged with a label (pneumothorax or no pneumothorax) and the CNN learning which pixel patterns suggest pneumothorax. When fed with new untagged images, the CNN outputs a likelihood of the new image containing a pneumothorax or not.
- Feeding in raw retinal images tagged with the stage of age related macular degeneration and the CNN learning which pixel patterns suggest a particular stage. When fed with new untagged images, the CNN outputs a likelihood of the new image containing a specific stage of age related macular degeneration.
- Feeding in optical coherence tomography scans tagged with a management decision (urgent referral, semi urgent referral, routine referral, observation). When fed with new untagged images, the CNN outputs a likelihood of the most appropriate management decision.

with disagreements resolved by consensus or a third reviewer.

Adherence to reporting standards and risk of bias

We assessed reporting quality of non-randomised studies against a modified version of the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement.[14] This statement aims to improve the transparent reporting of prediction modelling studies of all types and in all medical settings.[15] The TRIPOD statement consists of a 22 item checklist (37 total points when all subitems are included), but we considered some items to be less relevant to deep learning studies (eg, points that related to predictor variables). Deep learning algorithms can consider multiple predictors; however, in the cases we assessed, the only predictors (almost exclusively) were the individual pixels of the image. The algorithm did not typically receive information on characteristics such as patient age, sex, and medical history. Therefore, we used a modified list of 29 total points (see appendix 2). The aim was to assess whether studies broadly conformed to reporting recommendations included in TRIPOD, and not the detailed granularity required for a full assessment of adherence.[16]

We assessed risk of bias for non-randomised studies by applying PROBAST (prediction model risk of bias assessment tool).[17] PROBAST contains 20 signalling questions from four domains (participants, predictors, outcomes, and analysis) to allow assessment of the risk of bias in predictive modelling studies.[18] We did not assess applicability (because no specific therapeutic question existed for this systematic review) or predictor variables (these are less relevant in deep learning studies on medical imaging; see appendix 2).

We assessed the broad level reporting of randomised studies against the CONSORT (consolidated standards of reporting trials) statement. Risk of bias was evaluated by applying the Cochrane risk of bias tool.[11 19]

Data synthesis

We intentionally planned not to conduct formal quantitative syntheses because of the probable heterogeneity of specialties and outcomes.

Patient and public involvement

Patients were not involved in any aspect of the study design, conduct or in the development of the research question or outcome measures.

Results

Study selection

Our electronic search, which was last updated on 17 June 2019, retrieved 8302 records (7334 study records and 968 trial registrations; see fig 1). Of the 7334 study records, we assessed 140 full text articles; 59 were excluded, which left 81 non-randomised studies for analysis. Of the 968 trial registrations, we assessed 96 in full; 86 were excluded, which left 10 trial registrations that related to deep learning.

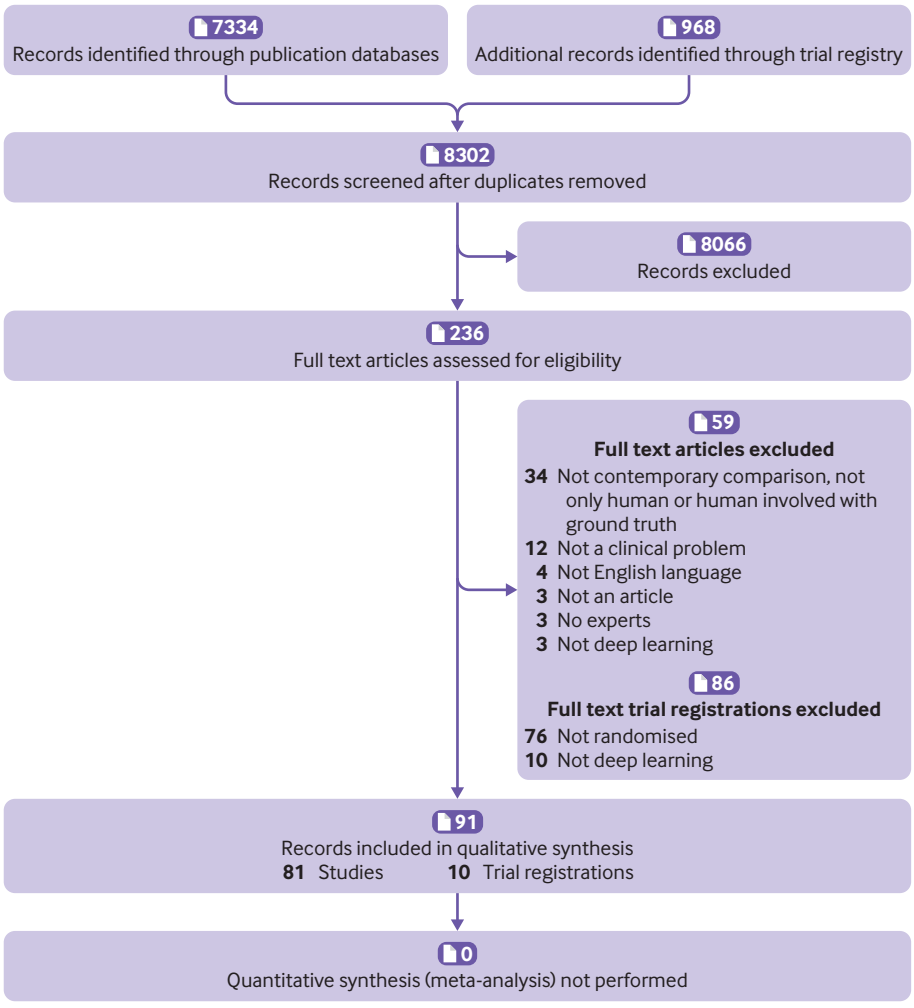

Fig 1 | PRISMA (preferred reporting items for systematic reviews and meta-analyses) flowchart of study records

Randomised clinical trials

Table 1 summarises the 10 trial registrations. Eight related to gastroenterology, one to ophthalmology, and one to radiology. Eight were from China, one was from the United States, and one from Taiwan. Two trials have completed and published their results (both in 2019), three are recruiting, and five are not yet recruiting.

The first completed trial enrolled 350 paediatric patients who attended ophthalmology clinics in China. These patients underwent cataract assessment with or without an AI platform (using deep learning) to diagnose and provide a treatment recommendation (surgery or follow-up).[20] The authors found that accuracy (defined as proportion of true results) of cataract diagnosis and treatment recommendation with AI were 87% (sensitivity 90%, specificity 86%) and 71% (sensitivity 87%, specificity 44%), respectively. These results were significantly lower than accuracy of diagnosis (99%, sensitivity 98%, specificity 99.6%) and treatment recommendation (97%, sensitivity 95%, specificity 100%) by senior consultants (P<0.001 for both); and also lower than the results for the same AI when tested in a non-randomised clinical trial setting (98% and 93%, respectively). The mean time for receiving a diagnosis with the AI platform was faster than diagnosis by consultants (2.8 *v* 8.5 minutes, P<0.001). The authors suggested that this might explain why patients were more satisfied with AI (mean satisfaction score 3.47 *v* 3.38, P=0.007). Risk of bias was low in all domains except for blinding of participants and personnel. The reporting showed high adherence (31 of 37 items, 84%) to the CONSORT checklist (which was included with the manuscript).

The second completed trial enrolled 1058 patients who underwent a colonoscopy with or without the assistance of a real time automatic polyp detection system, which provided simultaneous visual and sound alerts when it found a polyp.[21] The authors reported that the detection system resulted in a significant increase in the adenoma detection rate (29% *v* 20%, P<0.001), and an increase in the number of hyperplastic polyps identified (114 *v* 52, P<0.001). Risk of bias was low in all domains except for blinding of participants, personnel, and outcome assessors. One of the other trial registrations belongs to the same author group. These authors are performing a

double blind randomised clinical trial with sham AI to overcome the blinding issue in the previous study. The reporting showed high adherence (30 of 37 items, 81%) to the CONSORT checklist (though the CONSORT checklist itself was not included or referenced by the manuscript).

Non-randomised studies

General characteristics

Table 2 and table 3 summarise the basic characteristics of the 81 non-randomised studies. Nine of 81 (11%) non-randomised studies were prospective, but only six of these nine were tested in a real world clinical environment. The US and Asia accounted for 82% of studies, with the top four countries as follows: US (24/81, 30%), China (14/81, 17%), South Korea (12/81, 15%), and Japan (9/81, 11%). The top five specialties were radiology (36/81, 44%), ophthalmology (17/81, 21%), dermatology (9/81, 11%), gastroenterology (5/81, 6%), and histopathology (5/81, 6%). Eighteen (22%) studies compared how long a task took in AI and human arms in addition to accuracy or performance metrics. Funding was predominantly academic (47/81, 58%) as opposed to commercial (9/81, 11%) or mixed (1/81, 1%). Twelve studies stated they had no funding and another 12 did not report on funding. A detailed table with further information on the 81 studies is included as an online supplementary file.

In 77 of 81 studies, a specific comment was included in the abstract about the comparison between AI and clinician performance. AI was described as superior in 23 (30%), comparable or better in 13 (17%), comparable in 25 (32%), able to help a clinician perform better in 14 (18%), and not superior in two (3%). Only nine studies added a caveat in the abstract that further prospective trials were required (this was missing in all 23 studies that reported AI was superior to clinician performance). Even in the discussion section of the paper, a call for prospective studies (or trials in the case of existing prospective work) was only made in 31 of 81 (38%) studies. Seven of 81 (9%) studies claimed in the discussion that the algorithm could now be used in clinical practice despite only two of the seven having been tested prospectively in a real world setting. Concerning reproducibility, data were public and available in only four studies (5%). Code (for preprocessing of data and modelling) was available in only six studies (7%). Both raw labelled data and code were available in only one study.[22]

Methods and risk of bias

Most studies developed and validated a model (63/81, 78%) compared with development only by using validation through resampling (9/81, 11%) or validation only (9/81, 11%). When validation occurred in a separate dataset, this dataset was from a different geographical region in 19 of 35 (54%) studies, from a different time period in 11 of 35 (31%), and a combination of both in five of 35 (14%). In studies that did not use a separate dataset for validation, the most common method of internal validation was split sample (29/37) followed by cross validation (15/37), and then bootstrapping (6/37); some studies used more than one method (box 2). Sample size calculations were reported in 14 of 81 (17%) studies. Dataset sizes were as follows (when reported): training, median 2678 (interquartile range 704-21 362); validation, 600 (200-1359); and test, 337 (144-891). The median event rate for development, validation, and test sets was 42%, 44%, and 44%, respectively, when a binary outcome was assessed (n=62) as opposed to a multiclass classification (n=19). Forty one of 81 studies used data augmentation (eg, flipping and inverting images) to increase the dataset size.

The human comparator group was generally small and included a median of five clinicians (interquartile range 3-13, range 1-157), of which a median of four were experts (interquartile range 2-9, range 1-91). The number of participating non-experts varied from 0 to 94 (median 1, interquartile range 0-3). Experts were used exclusively in 36 of 81 studies, but in the 45 studies that included non-experts, 41 had separate performance data available which were exclusive to the expert group. In most studies, every human (expert or non-expert) rated the test dataset independently (blinded to all other clinical information except the image in 33/81 studies). The volume and granularity of the separate data for experts varied considerably among studies, with some reporting individual performance metrics for each human (usually in supplementary appendices).

The overall risk of bias assessed using PROBAST led to 58 of 81 (72%) studies being classified as high risk (fig 2); the analysis domain was most commonly rated to be at high risk of bias (as opposed to participant or outcome ascertainment domains). Major deficiencies in the analysis domain related to PROBAST items 4.1 (were there a reasonable number of participants?), 4.3 (were all enrolled participants included in the analysis?), 4.7 (were relevant model performance measures evaluated appropriately?), and 4.8 (were model overfitting and optimism in model performance accounted for?).

Adherence to reporting standards

Adherence to reporting standards was poor (<50% adherence) for 12 of 29 TRIPOD items (see fig 3). Overall, publications adhered to between 24% and 90% of the TRIPOD items: median 62% (interquartile range 45-69%). Eight TRIPOD items were reported in 90% or more of the 81 studies, and five items in less than 30% (fig 3). A flowchart for the flow of patients or data through the study was only present in 25 of 81 (31%) studies. We also looked for reporting of the hardware that was used for developing or validating the algorithm, although this was not specifically requested in the TRIPOD statement. Only 29 of 81 (36%) studies reported this information and in most cases (n=18) it related only to the graphics processing unit rather than providing full details (eg, random access memory, central processing unit speed, configuration settings).

Table 1 | Randomised trial registrations of deep learning algorithms

Trial registration	Title	Status	Record last updated	Country	Specialty	Planned sample size	Intervention	Control	Blinding	Primary outcome	Anticipated completion
ChiCTR-DDD-17012221	A colorectal polyps auto-detection system based on deep learning to increase polyp detection rate: a prospective clinical study	Completed, published	16 July 2018	China	Gastroenterology	1000	AI assisted colonoscopy	Standard colonoscopy	None	Polyp detection rate and adenoma detection rate	28 February 2018
NCT03240848	Comparison of artificial intelligent clinic and normal clinic	Completed, published	30 July 2018	China	Ophthalmology	350	AI assisted clinic	Normal clinic	Double (investigator and outcomes assessor)	Accuracy for congenital cataracts	25 May 2018
NCT03706534	Breast ultrasound image reviewed with assistance of deep learning algorithms	Recruiting	17 October 2018	US	Radiology	300	Computer aided detection system	Manual ultrasound imaging review	Double (participant and investigator)	Concordance rate	31 July 2019
NCT03840590	Adenoma detection rate using AI system in China	Not yet recruiting	15 February 2019	China	Gastroenterology	800	CSK AI system assisted colonoscopy	Standard colonoscopy	None	Adenoma detection rate	1 March 2020
NCT03842059	Computer-aided detection for colonoscopy	Not yet recruiting	15 February 2019	Taiwan	Gastroenterology	1000	Computer aided detection	Standard colonoscopy	Double (participant, care provider)	Adenoma detection rate	31 December 2021
ChiCTR1800017675	The impact of a computer aided diagnosis system based on deep learning on increasing polyp detection rate during colonoscopy, a prospective double blind study	Not yet recruiting	21 February 2019	China	Gastroenterology	1010	AI assisted colonoscopy	Standard colonoscopy	Double	Polyp detection rate and adenoma detection rate	31 January 2019
ChiCTR1900021984	A multicenter randomised controlled study for evaluating the effectiveness of artificial intelligence in improving colonoscopy quality	Recruiting	19 March 2019	China	Gastroenterology	1320	EndoAngel assisted colonoscopy	Colonoscopy	Double (participants and evaluators)	Polyp detection rate	31 December 2020
NCT03908645	Development and validation of a deep learning algorithm for bowel preparation quality scoring	Not yet recruiting	9 April 2019	China	Gastroenterology	100	AI assisted scoring group	Conventional human scoring group	Single (outcome assessor)	Adequate bowel preparation	15 April 2020
NCT03883035	Quality measurement of esophago-gastroduodenoscopy using deep learning models	Recruiting	17 April 2019	China	Gastroenterology	559	DCNN model assisted EGD	Conventional EGD	Double (participant, care provider)	Detection of upper gastrointestinal lesions	20 May 2020
ChiCTR1900023282	Prospective clinical study for artificial intelligence platform for lymph node pathology detection of gastric cancer	Not yet recruiting	20 May 2019	China	Gastroenterology	60	Pathological diagnosis of artificial intelligence	Traditional pathological diagnosis	Not stated	Clinical prognosis	31 August 2021

AI=artificial intelligence; CSK=commonsense knowledge; DCNN=deep convolutional neural network; EGD=esophagogastroduodenoscopy.

doi: 10.1136/bmj.m689 | BMJ 2020;368:m689 | the bmj

Table 2 | Characteristics of non-randomised studies

Lead author	Year	Country	Study type	Specialty	Disease	Outcome	Caveat in discussion*	Suggestion in discussion†
Abramoff	2018	US	Prospective real world	Ophthalmology	Diabetic retinopathy	More than mild diabetic retinopathy	No	Yes
Arbabshirani	2018	US	Prospective real world	Radiology	Intracranial haemorrhage	Haemorrhage	Yes	No
Arji	2018	Japan	Retrospective	Radiology	Oral cancer	Cervical lymph node metastases	No	No
Becker	2017	Switzerland	Retrospective	Radiology	Breast cancer	BI-RADS category 5	Yes	No
Becker	2018	Switzerland	Retrospective	Radiology	Breast cancer	BI-RADS category 5	Yes	No
Bien	2018	US	Retrospective	Radiology	Knee injuries	Abnormality on MRI	No	No
Brinker	2019	Germany	Retrospective	Dermatology	Skin cancer	Melanoma	No	No
Brinker	2019	Germany	Retrospective	Dermatology	Skin cancer	Melanoma	Yes	No
Brown	2018	US	Retrospective	Ophthalmology	Retinopathy of prematurity	Plus disease	No	No
Burlina	2018	US	Retrospective	Ophthalmology	Macular degeneration	ARMD stage	No	No
Burlina	2017	US	Retrospective	Ophthalmology	Macular degeneration	Intermediate or advanced stage ARMD	No	No
Burlina	2017	US	Retrospective	Ophthalmology	Macular degeneration	ARMD stage	No	No
Bychov	2018	Finland	Retrospective	Histopathology	Colorectal cancer	Low or high risk for 5 year survival	No	No
Byra	2018	US	Retrospective	Radiology	Breast cancer	BI-RADS category 4 or more	No	No
Cha	2018	US	Retrospective	Radiology	Bladder cancer	T0 status post chemotherapy	No	No
Cha	2019	South Korea	Retrospective	Radiology	Lung cancer	Nodule operability	Yes	No
Chee	2019	South Korea	Retrospective	Radiology	Osteonecrosis of the femoral head	Stage of osteonecrosis	Yes	No
Chen	2018	Taiwan	Prospective	Gastroenterology	Colorectal cancer	Neoplastic polyp	No	No
Choi	2018	South Korea	Retrospective	Radiology	Liver fibrosis	Fibrosis stage	No	No
Choi	2019	South Korea	Retrospective	Radiology	Breast cancer	Malignancy	No	No
Chung	2018	South Korea	Retrospective	Orthopaedics	Humerus fractures	Proximal humerus fracture	Yes	No
Ciompi	2017	Netherlands/Italy	Retrospective	Radiology	Lung cancer	Nodule type	No	No
Ciritsis	2019	Switzerland	Retrospective	Radiology	Breast cancer	BI-RADS stage	No	No
De Fauw	2018	UK	Retrospective	Ophthalmology	Retinopathy	Diagnosis and referral decision	Yes	No
Ehtesham Bejnordii	2017	Netherlands	Retrospective	Histopathology	Breast cancer	Metastases	Yes	No
Esteva	2017	US	Retrospective	Dermatology	Skin cancer	Lesion type	Yes	No
Fujioka	2019	Japan	Retrospective	Radiology	Breast cancer	BI-RADS malignancy	No	No
Fujisawa	2018	Japan	Retrospective	Dermatology	Skin cancer	Malignancy classification	Yes	No
Gan	2019	China	Retrospective	Orthopaedics	Wrist fractures	Fracture	Yes	No
Gulshan	2019	India	Prospective real world	Ophthalmology	Retinopathy	Moderate or worse diabetic retinopathy or referable macula oedema	Yes	No
Haenssle	2018	Germany	Retrospective	Dermatology	Skin cancer	Malignancy classification and management decision	Yes	No
Hamm	2019	US	Retrospective	Radiology	Liver cancer	LI-RADS category	No	No
Han	2018	South Korea	Retrospective	Dermatology	Skin cancer	Cancer type	No	No
Han	2018	South Korea	Retrospective	Dermatology	Onchomycosis	Onchomycosis diagnosis	No	No
Hannun	2019	US	Retrospective	Cardiology	Arrhythmia	Arrhythmia classification	No	No
He	2019	China	Retrospective	Radiology	Bone cancer	Recurrence of giant cell tumour	No	No
Hwang	2019	Taiwan	Retrospective	Ophthalmology	Macular degeneration	Classification and type of ARMD	No	Yes
Hwang	2018	South Korea	Retrospective	Radiology	Tuberculosis	TB presence	Yes	No
Hwang	2019	South Korea	Retrospective	Radiology	Pulmonary pathology	Abnormal chest radiograph	Yes	No
Kim	2018	South Korea	Retrospective	Radiology	Sinusitis	Maxillary sinusitis label	No	No
Kise	2019	Japan	Retrospective	Radiology	Sjogren's syndrome	Sjogren's syndrome presence	No	No
Kooi	2017	Netherlands	Retrospective	Radiology	Breast cancer	Classification of mammogram	No	No
Krause	2018	US	Retrospective	Ophthalmology	Diabetic retinopathy	Diabetic retinopathy stage	No	No
Kuo	2019	Taiwan	Retrospective	Nephrology	Chronic kidney disease	eGFR<60 mL/min/1.73m^2	No	Yes
Lee	2018	US	Prospective	Radiology	Intracranial haemorrhage	Haemorrhage	Yes	No
Li	2018	China	Prospective	oncology	Nasopharyngeal cancer	Malignancy	No	Yes
Li	2018	China	Retrospective	Ophthalmology	Glaucoma	Glaucoma	No	No
Li	2018	China	Retrospective	Radiology	Thyroid cancer	Malignancy	Yes	No

ARMD=age related macular degeneration; BI-RADS=breast imaging reporting and data system; eGFR=estimated glomerular filtration rate; LI-RADS=liver imaging reporting and data system; MRI=magnetic resonance imaging; TB=tuberculosis.
*Caveat mentioned in discussion about need for further prospective work or trials.
†Suggestion in discussion that algorithm can now be used clinically.

Table 3 | Characteristics of non-randomised studies

Lead author	Year	Country	Study type	Specialty	Disease	Outcome	Caveat in discussion*	Suggestion in discussion†
Long	2017	China	Prospective real world	Ophthalmology	Congenital cataracts	Detection of congenital cataracts	No	No
Lu	2018	China	Retrospective	Ophthalmology	Macular pathologies	Classification of macular pathology	No	No
Marchetti	2017	US	Retrospective	Dermatology	Skin cancer	Malignancy (melanoma)	Yes	No
Matsuba	2018	Japan	Retrospective	Ophthalmology	Macular degeneration	Wet AMD	No	No
Mori	2018	Japan	Prospective real world	Gastroenterology	Polyps	Neoplastic polyp	Yes	Yes
Nagpal	2019	US	Retrospective	Histopathology	Prostate cancer	Gleason score	No	No
Nakagawa	2019	Japan	Retrospective	Gastroenterology	Oesophageal cancer	Cancer invasion depth stage SM2/3	No	No
Nam	2018	South Korea	Retrospective	Radiology	Pulmonary nodules	Classification and localisation of nodule	Yes	No
Nirschl	2018	US	Retrospective	Histopathology	Heart failure	Heart failure (pathologically)	No	Yes
Olczak	2017	Sweden	Retrospective	Orthopaedics	Fractures	Fracture	No	Yes
Park	2019	US	Retrospective	Radiology	Cerebral aneurysm	Aneurysm presence	Yes	No
Poedjiastoeti	2018	Thailand	Retrospective	Oncology	Jaw tumours	Malignancy	No	No
Rajpurkar	2018	US	Retrospective	Radiology	Pulmonary pathology	Classification of chest radiograph pathology	Yes	No
Raumviboonsuk	2019	Thailand	Prospective real world	Ophthalmology	Diabetic retinopathy	Moderate or worse diabetic retinopathy	Yes	No
Rodriguez-Ruiz	2018	Netherlands	Retrospective	Radiology	Breast cancer	Classification of mammogram	Yes	No
Sayres	2019	US	Retrospective	Ophthalmology	Diabetic retinopathy	Moderate or worse non-proliferative diabetic retinopathy	No	No
Shichijo	2017	Japan	Retrospective	Gastroenterology	Gastritis	*Helicobacter pylori* gastritis	No	No
Singh	2018	US	Retrospective	Radiology	Pulmonary pathology	Chest radiograph abnormality	No	No
Steiner	2018	US	Retrospective	Histopathology	Breast cancer	Metastases	Yes	No
Ting	2017	Singapore	Retrospective	Ophthalmology	Retinopathy, glaucoma, macular degeneration	Referable pathology for retinopathy, glaucoma, macular degeneration	Yes	No
Urakawa	2019	Japan	Retrospective	Orthopaedics	Hip fractures	Intertrochanteric hip fracture	No	No
van Grinsven	2016	Netherlands	Retrospective	Ophthalmology	Fundal haemorrhage	Fundal haemorrhage	No	No
Walsh	2018	UK/Italy	Retrospective	Radiology	Fibrotic lung disease	Fibrotic lung disease	No	No
Wang	2019	China	Retrospective	Radiology	Thyroid nodule	Nodule presence	Yes	No
Wang	2018	China	Retrospective	Radiology	Lung cancer	Invasive or preinvasive adenocarcinoma nodule	No	No
Wu	2019	US	Retrospective	Radiology	Bladder cancer	T0 response to chemotherapy	No	No
Xue	2017	China	Retrospective	Orthopaedics	Hip osteoarthritis	Radiograph presence of hip osteoarthritis	No	No
Ye	2019	China	Retrospective	Radiology	Intracranial haemorrhage	Presence of intracranial haemorrhage	Yes	No
Yu	2018	South Korea	Retrospective	Dermatology	Skin cancer	Malignancy (melanoma)	No	No
Zhang	2019	China	Retrospective	Radiology	Pulmonary nodules	Presence of a malignant nodule	Yes	No
Zhao	2018	China	Retrospective	Radiology	Lung cancer	Classification of nodule invasiveness	No	No
Zhu	2019	China	Retrospective	Gastroenterology	Gastric cancer	Tumour invasion depth (deeper than SM1)	No	No
Zucker	2019	US	Retrospective	Radiology	Cystic fibrosis	Brasfield score	Yes	No

AMD=age related macular degeneration.
*Caveat mentioned in discussion about need for further prospective work or trials.
†Suggestion in discussion that algorithm can now be used clinically.

doi: 10.1136/bmj.m689 | *BMJ* 2020;368:m689 | the bmj

Box 2: Specific terms

- Internal validation: evaluation of model performance with data used in development process
- External validation: evaluation of model performance with separate data not used in development process
- Cross validation: internal validation approach in which data are randomly split into n equally sized groups; the model is developed in n−1 of n groups, and performance evaluated in the remaining group with the whole process repeated n times; model performance is taken as average over n iterations
- Bootstrapping: internal validation approach similar to cross validation but relying on random sampling with replacement; each sample is the same size as model development dataset
- Split sample: internal validation approach in which the available development dataset is divided into two datasets: one to develop the model and the other to validate the model; division can be random or non-random.

Discussion

We have conducted an appraisal of the methods, adherence to reporting standards, risk of bias, and claims of deep learning studies that compare diagnostic AI performance with human clinicians. The rapidly advancing nature and commercial drive of this field has created pressure to introduce AI algorithms into clinical practice as quickly as possible. The potential consequences for patients of this implementation without a rigorous evidence base make our findings timely and should guide efforts to improve the design, reporting, transparency, and nuanced conclusions of deep learning studies.[23 24]

Principal findings

Five key findings were established from our review. Firstly, we found few relevant randomised clinical trials (ongoing or completed) of deep learning in medical imaging. While time is required to move from development to validation to prospective feasibility testing before conducting a trial, this means that claims about performance against clinicians should be tempered accordingly. However, deep learning only became mainstream in 2014, giving a lead time of approximately five years for testing within clinical environments, and prospective studies could take a minimum of one to two years to conduct. Therefore, it is reasonable to assume that many similar trials will be forthcoming over the next decade. We found only one randomised trial registered in the US despite at least 16 deep learning algorithms for medical imaging approved for marketing by the Food and Drug Administration (FDA). These algorithms cover a range of fields from radiology to ophthalmology and cardiology.[2 25]

Secondly, of the non-randomised studies, only nine were prospective and just six were tested in a real world clinical environment. Comparisons of AI performance against human clinicians are therefore difficult to evaluate given the artificial in silico context in which clinicians are being evaluated. In much the same way that surrogate endpoints do not always reflect clinical benefit,[26] a higher area under the curve might not lead to clinical benefit and could even have unintended adverse effects. Such effects could include an unacceptably high false positive rate, which is not apparent from an in silico evaluation. Yet it is typically retrospective studies that are usually cited in FDA approval notices for marketing of algorithms. Currently, the FDA do not mandate peer reviewed publication of these studies; instead internal review alone is performed.[27 28] However, the FDA has recognised and acknowledged that their traditional paradigm of medical device regulation was not designed for adaptive AI and machine learning technologies. Non-inferior AI (rather than superior) performance that allows for a lower burden on clinician workflow (that is, being quicker with similar accuracy) might warrant further investigation. However, less than a quarter of studies reported time taken for task completion in both the AI and human groups. Ensuring fair comparison between AI and clinicians is arguably done best in a randomised clinical trial (or at the very least prospective) setting. However, it should be noted that prospective testing is not necessary to actually develop the model in the first place. Even in a randomised clinical trial setting, ensuring that functional robustness tests are present is crucial. For example, does the algorithm produce the correct decision for normal anatomical variants and is the decision independent of the camera or imaging software used?

Thirdly, limited availability of datasets and code makes it difficult to assess the reproducibility of deep learning research. Descriptions of the hardware used, when present, were also brief and this vagueness might affect external validity and implementation. Reproducible research has become a pressing issue across many scientific disciplines and efforts to encourage data and code sharing are crucial.[29-31] Even when commercial concerns exist about intellectual property, strong arguments exist for ensuring that algorithms are non-proprietary and available for scrutiny.[32] Commercial companies could collaborate with non-profit third parties for independent prospective validation.

Fourthly, the number of humans in the comparator group was typically small with a median of only four experts. There can be wide intra and inter case variation even between expert clinicians. Therefore, an appropriately large human sample for comparison is essential for ensuring reliability. Inclusion of non-experts can dilute the average human performance and potentially make the AI algorithm look better than it otherwise might. If the algorithm is designed specifically to aid performance of more junior clinicians or non-specialists rather than experts, then this should be made clear.

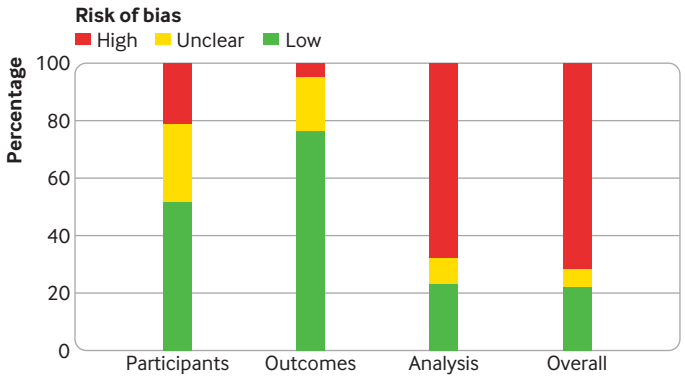

Fig 2 | PROBAST (prediction model risk of bias assessment tool) risk of bias assessment for non-randomised studies

Fifthly, descriptive phrases that suggested at least comparable (or better) diagnostic performance of an algorithm to a clinician were found in most abstracts, despite studies having overt limitations in design, reporting, transparency, and risk of bias. Caveats about the need for further prospective testing were rarely mentioned in the abstract (and not at all in the 23 studies that claimed superior performance to a clinician). Accepting that abstracts are usually word limited, even in the discussion sections of the main text, nearly two thirds of studies failed to make an explicit recommendation for further prospective studies or trials. One retrospective study gave a website address in the abstract for patients to upload their eye scans and use the algorithm themselves.[33] Overpromising language leaves studies vulnerable to being misinterpreted by the media and the public. Although it is clearly beyond the power of authors to control how the media and public interpret their findings, judicious and responsible use of language in studies and press releases that factor in the strength and quality of the evidence can help.[34] This issue is especially concerning given the findings from new research that suggests patients are more likely to consider a treatment beneficial when news stories are reported with spin, and that false news spreads much faster online than true news.[35 36]

Policy implications

The impetus for guiding best practice has gathered pace in the last year with the publication of a report that proposes a framework for developing transparent, replicable, ethical, and effective research in healthcare AI (AI-TREE).[37] This endeavour is led by a multidisciplinary team of clinicians, methodologists, statisticians, data scientists, and healthcare policy makers. The guiding questions of this framework will probably feed into the creation of more specific reporting standards such as a TRIPOD extension for machine learning studies.[38] Key to the success of these efforts will be high visibility to researchers and possibly some degree of enforcement by journals in a similar vein to preregistering randomised trials and reporting them according to the CONSORT statement.[11 39] Enthusiasm exists to speed up the process by which medical devices that feature AI are approved for marketing.[40 41] Better design and more transparent reporting should be seen eventually as a facilitator of the innovation, validation, and translation process, and could help avoid hype.

Study limitations

Our findings must be considered in light of several limitations. Firstly, although comprehensive, our search might have missed some studies that could have been included. Secondly, the guidelines that we used to assess non-randomised studies (TRIPOD and PROBAST) were designed for conventional prediction modelling studies, and so the adherence levels we found should be interpreted in this context. Thirdly, we focused specifically on deep learning for diagnostic medical imaging. Therefore, it might not be appropriate

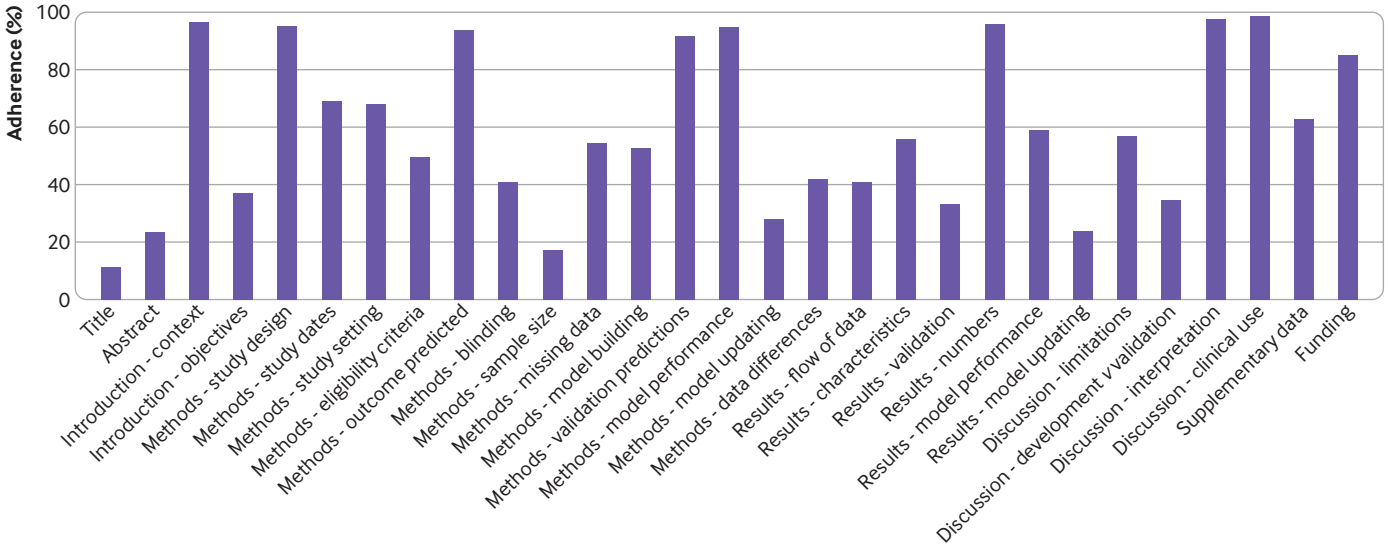

Fig 3 | Completeness of reporting of individual TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) items for non-randomised studies

to generalise our findings to other types of AI, such as conventional machine learning (eg, an artificial neural network based mortality prediction model that uses electronic health record data). Similar issues could exist in many other types of AI paper, however we cannot definitively make this claim from our findings because we only assessed medical imaging studies. Moreover, nomenclature in the field is sometimes used in non-standardised ways, and thus some potentially eligible studies might have been presented with terminology that did not lead to them being captured with our search strategy. Fourthly, risk of bias entails some subjective judgment and people with different experiences of AI performance could have varying perceptions.

Conclusions

Deep learning AI is an innovative and fast moving field with the potential to improve clinical outcomes. Financial investment is pouring in, global media coverage is widespread, and in some cases algorithms are already at marketing and public adoption stage. However, at present, many arguably exaggerated claims exist about equivalence with or superiority over clinicians, which presents a risk for patient safety and population health at the societal level, with AI algorithms applied in some cases to millions of patients. Overpromising language could mean that some studies might inadvertently mislead the media and the public, and potentially lead to the provision of inappropriate care that does not align with patients' best interests. The development of a higher quality and more transparently reported evidence base moving forward will help to avoid hype, diminish research waste, and protect patients.

AUTHOR AFFILIATIONS

[1]Division of Anaesthetics, Pain Medicine and Intensive Care, Department of Surgery and Cancer, Imperial College London, UK

[2]Institute of Cardiovascular Science, University College London, UK

[3]Cera Care, London, UK

[4]Centre for Perioperative and Critical Care Research, Imperial College Healthcare NHS Trust, London, UK

[5]Department of Bioengineering, Imperial College London, London, UK

[6]Hardian Health, London, UK

[7]Scripps Research Translational Institute, La Jolla, California, USA

[8]Departments of Medicine, of Health Research and Policy, of Biomedical Data Sciences, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

[9]Centre for Statistics in Medicine, University of Oxford, Oxford, UK

[10]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Trust, Oxford, UK

We thank Dina Radenkovic for assistance with sorting through search results and selection of includable studies. We thank the BMJ editors and peer reviewers for extensive comments and suggestions which have been incorporated into the manuscript.

Contributors: MN and MM conceived the study. MN, YC, and CAL executed the search and extracted data. MN performed the initial analysis of data, with all authors contributing to interpretation of data. JPAI contributed to amendments on the protocol. All authors contributed to critical revision of the manuscript for important intellectual content and approved the final version. MN is the study guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: There is no specific funding for this study. MN and YC are supported by National Institute for Health Research (NIHR) academic clinical fellowships. ACG is funded by a UK NIHR research professor award (RP-2015-06-018). MN and ACG are both supported by the NIHR Imperial Biomedical Research Centre. The Meta-Research Innovation Center at Stanford (METRICS) has been funded by a grant from the Laura and John Arnold Foundation. GSC is supported by the NIHR Oxford Biomedical Research Centre and Cancer Research UK (grant C49297/A27294).

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; CAL worked as clinical data science and technology lead for Cera, a technology enabled homecare provider; ACG reports that outside of this work he has received speaker fees from Orion Corporation Orion Pharma and Amomed Pharma, has consulted for Ferring Pharmaceuticals, Tenax Therapeutics, Baxter Healthcare, Bristol-Myers Squibb and GSK, and received grant support from Orion Corporation Orion Pharma, Tenax Therapeutics, and HCA International with funds paid to his institution; HH was previously clinical director of Kheiron Medical Technologies and is now director at Hardian Health; EJT is on the scientific advisory board of Verily, Tempus Laboratories, Myokardia, and Voxel Cloud, the board of directors of Dexcoman, and is an advisor to Guardant Health, Blue Cross Blue Shield Association, and Walgreens; MM is a cofounder of Cera, a technology enabled homecare provider, board member of the NHS Innovation Accelerator, and senior advisor to Bain and Co.

Ethical approval: Not required.

Data sharing: Raw data are available on request from the corresponding author.

The lead author and manuscript's guarantor (MN) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: We plan to use social media to help disseminate the findings from this research as well as engaging with patient groups. The timing of this dissemination will begin with the publication of this article and continue during early 2020.

1 Carson E. IBM Watson Health computes a pair of new solutions to improve healthcare data and security. 2015. https://www.techrepublic.com/article/ibm-watson-health-computes-a-pair-of-new-solutions-to-improve-healthcare-data-and-security/.

2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56. doi:10.1038/s41591-018-0300-7

3 ReportLinker. Artificial intelligence in healthcare market by offering, technology, end-use application, end user and geography – global forecast to 2025. 2018. https://www.reportlinker.com/p04897122/Artificial-Intelligence-in-Healthcare-Market-by-Offering-Technology-Application-End-User-Industry-and-Geography-Global-Forecast-to.html.

4 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436-44. doi:10.1038/nature14539

5 Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019;25:24-9. doi:10.1038/s41591-018-0316-z

6 NCBI. PubMed search for deep learning. 2019. https://www.ncbi.nlm.nih.gov/pubmed/?term=deep+learning+or+%22Deep+Learning%22%5BMesh%5D.

7 Murphy M. Google says its AI can spot lung cancer a year before doctors. 2019. https://www.telegraph.co.uk/technology/2019/05/07/google-says-ai-can-spot-lung-cancer-year-doctors/.

8 Price E. AI Is better at diagnosing skin cancer than your doctor, study finds. 2018. https://fortune.com/2018/05/30/ai-skin-cancer-diagnosis/.

9 Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res* 2018;2:11. doi:10.1186/s41512-018-0033-6

10 Psaty BM, Furberg CD. COX-2 inhibitors – lessons in drug safety. *N Engl J Med* 2005;352:1133-5. doi:10.1056/NEJMe058042

11 Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c332. doi:10.1136/bmj.c332

12 Wallace E, Smith SM, Perera-Salazar R, et al, International Diagnostic and Prognosis Prediction (IDAPP) group. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC Med Inform Decis Mak* 2011;11:62. doi:10.1186/1472-6947-11-62
13 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535. doi:10.1136/bmj.b2535
14 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594. doi:10.1136/bmj.g7594
15 Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. doi:10.7326/M14-0698
16 Heus P, Damen JAAG, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open* 2019;9:e025611. doi:10.1136/bmjopen-2018-025611
17 Wolff RF, Moons KGM, Riley RD, et al, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51-8. doi:10.7326/M18-1376
18 Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019;170:W1-33. doi:10.7326/M18-1377
19 Higgins JP, Altman DG, Gøtzsche PC, et al, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928. doi:10.1136/bmj.d5928
20 Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* 2019;9:52-9. doi:10.1016/j.eclinm.2019.03.001
21 Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68:1813-9. doi:10.1136/gutjnl-2018-317500
22 Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS One* 2018;13:e0191493. doi:10.1371/journal.pone.0191493
23 Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383:267-76. doi:10.1016/S0140-6736(13)62228-X
24 Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383:166-75. doi:10.1016/S0140-6736(13)62227-8
25 Carfagno J. 5 FDA Approved Uses of AI in Healthcare. 2019. https://www.docwirenews.com/docwire-pick/future-of-medicine-picks/fda-approved-uses-of-ai-in-healthcare/.
26 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled?*Ann Intern Med* 1996;125:605-13. doi:10.7326/0003-4819-125-7-199610010-00011
27 Section FDA. 510(k) premarket notification of intent to market the device. 2018. https://www.accessdata.fda.gov/cdrh_docs/pdf18/K180647.pdf.
28 FDA. Artificial Intelligence and Machine Learning in Software as a Medical Device. 2019. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
29 Camerer CF, Dreber A, Holzmeister F, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav* 2018;2:637-44. doi:10.1038/s41562-018-0399-z
30 Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. *JAMA* 2014;312:1024-32. doi:10.1001/jama.2014.9646
31 Wallach JD, Boyack KW, Ioannidis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLoS Biol* 2018;16:e2006930. doi:10.1371/journal.pbio.2006930
32 Van Calster B, Steyerberg EW, Collins GS. Artificial Intelligence Algorithms for Medical Prediction Should Be Nonproprietary and Readily Available. *JAMA Intern Med* 2019;179:731. doi:10.1001/jamainternmed.2019.0597
33 Hwang D-K, Hsu C-C, Chang K-J, et al. Artificial intelligence-based decision-making for age-related macular degeneration. *Theranostics* 2019;9:232-45. doi:10.7150/thno.28447
34 Sumner P, Vivian-Griffiths S, Boivin J, et al. Exaggerations and caveats in press releases and health-related science news. *PLoS One* 2016;11:e0168217. doi:10.1371/journal.pone.0168217
35 Boutron I, Haneef R, Yavchitz A, et al. Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit [correction in: *BMC Med* 2019;17:147]. *BMC Med* 2019;17:105. doi:10.1186/s12916-019-1330-9
36 Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science* 2018;359:1146-51. doi:10.1126/science.aap9559
37 Vollmer S, Mateen BA, Bohner G, et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. 2018. https://arxiv.org/abs/1812.10404.
38 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577-9. doi:10.1016/S0140-6736(19)30037-6
39 De Angelis C, Drazen JM, Frizelle FA, et al, International Committee of Medical Journal Editors. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;351:1250-1. doi:10.1056/NEJMe048225
40 Allen B. The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices. *J Am Coll Radiol* 2019;16:208-10. doi:10.1016/j.jacr.2018.09.007
41 Ratner M. FDA backs clinician-free AI imaging diagnostic tools. *Nat Biotechnol* 2018;36:673-4. doi:10.1038/nbt0818-673a

Web appendix: Supplementary appendices
Web appendix: Supplementary file