Artificial intelligence vs. clinicians – a systematic review of the design, reporting standards, and claims of deep learning studies in medical imaging

Journal:	*BMJ*
Manuscript ID	BMJ-2019-052733.R1
Article Type:	Research
BMJ Journal:	BMJ
Date Submitted by the Author:	01-Dec-2019
Complete List of Authors:	Nagendran, Myura; Imperial College London, Division of Anaesthetics, Pain Medicine and Intensive Care Chen, Yang; UCL, Institute of Cardiovascular Science Lovejoy, Christopher Andrew; Cera Care Gordon, Anthony; Imperial College London, Division of Anaesthetics, Pain Medicine and Intensive Care Komorowski, Matthieu; Imperial College London, Division of Anaesthetics, Pain Medicine and Intensive Care Harvey, Hugh; Hardian Health Topol, Eric; Scripps Research Translational Institute Ioannidis, John; Stanford University, Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy Collins, Gary; University of Oxford, Centre for Statistics in Medicine Maruthappu, Mahiben; Cera Care
Keywords:	artificial intelligence, deep learning, machine learning, reporting standards, systematic review

Artificial intelligence vs. clinicians – a systematic review of the design, reporting standards, and claims of deep learning studies in medical imaging

Myura Nagendran, FRCA[1]

Yang Chen, MRCP[2]

Christopher A Lovejoy, MB BChir[3]

Anthony C Gordon, MD[1,4]

Matthieu Komorowski, MD, PhD[1,5]

Hugh Harvey, MD, FRCR[6]

Eric J Topol, MD[7]

John P A Ioannidis, MD, DSc[8]

Gary S Collins, PhD[9,10]

Mahiben Maruthappu, BM BCh[3]

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, 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.

Version:	4.1
Date of version:	1 December 2019

Corresponding author: Dr Myura Nagendran MA, BM BCH, MRCP, FRCA, FFICM
NIHR Academic Clinical Fellow in Intensive Care Medicine
St Mary's Campus
Imperial College London
Praed Street, London
W2 1NY, UK

E-mail: myura.nagendran@imperial.ac.uk

Phone: +44 791 255 6717

Keywords: artificial intelligence; deep learning; reporting quality; methodology; systematic review

Word count: Abstract: **412** Manuscript: **3,955**

SUMMARY BOX

What is already known on this topic

- The volume of published research on deep learning, a branch of artificial intelligence (AI), for medical imaging is rapidly growing.
- In some cases, media headlines claiming superior performance to doctors have fuelled hype amongst the public and press for accelerated implementation.

What this study adds / what the problems are

- We found two published randomized clinical trials (RCTs) and 81 non-randomized studies (only six were prospectively tested in a real-world clinical setting).
- The overall risk of bias was high in more than two thirds of studies and adherence to reporting standards was suboptimal.
- Three quarters of studies stated in their abstract that the AI performance was at least comparable to (or better than) clinicians. Only one third stated that further prospective studies or trials were required.
- Limited availability of datasets and code make assessing the reproducibility of deep learning research in medical imaging difficult.
- The number of humans in the comparator group was typically small with a median of only four experts.

Suggestions for improvement

- More prospective studies with testing in a real-world clinical setting, ideally in an RCT.
- Lower risk of bias and greater adherence to reporting guidelines (this will be facilitated by the development of AI-specific reporting guidance and growing familiarity by clinical medical journals with deep learning research).
- More cautious language in the abstract when describing performance against clinicians and a clearer acknowledgement of the need for further prospective work (including potentially RCTs) before en-masse adoption.
- Better availability of datasets and code to enable reproducible research.
- Larger samples of expert clinicians in the human comparator group of studies.

ABSTRACT

Background and objectives

There is a rapidly growing volume of published research in medical imaging that examines deep learning, a branch of artificial intelligence (AI) in which an algorithm learns for itself the features of an image that are important for classification. In some cases, media headlines claiming superior performance to doctors have fuelled hype amongst the public and press for accelerated implementation. We aimed 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 against expert clinicians.

Design

Systematic review.

Data sources

Electronic database search of Medline, Embase, CENTRAL and the WHO trial registry up to June 2019 for studies that compared performance of a deep learning algorithm in medical imaging to a contemporary group of one or more expert clinicians. There was no limit placed on the aim or specific outcome measures used in these studies (absolute risk prediction [probability of disease] or classification [disease or not]). Adherence to reporting standards used CONSORT and TRIPOD for randomized and non-randomized studies respectively. Risk of bias assessment used the Cochrane risk of bias tool and PROBAST for randomized and non-randomized studies respectively.

Results

Trial registries revealed only ten records for deep learning RCTs. Two of these have been published (with low risk of bias [except for lack of blinding] and high adherence to reporting standards) while eight are ongoing. Of 81 non-RCTs 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 (IQR 2 to 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 58/81 studies and adherence to reporting standards was suboptimal (<50% adherence for 12/29 TRIPOD items). 61/81 studies stated in their abstract that the AI performance was at least comparable to (or better than) clinicians. Only 38% stated that further prospective studies or trials were required.

Conclusions

There are very few prospective deep learning studies in medical imaging and even fewer randomized trials. The majority of non-randomized studies are not prospective, and demonstrate substantive bias and deviation from existing reporting standards. Data and code availability is 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.

Registration

PROSPERO CRD42019123605

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 $2B in 2018 to $36B 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 deep learning and traditional machine learning is that once fed with raw data, deep learning algorithms develop their own representations needed for pattern recognition rather than requiring domain expertise to structure the data and design feature extractors.[5] In plain language, this means 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. Fields such as medical imaging have seen a growing interest in, and publication of, deep learning research.[6] In some cases, media headlines claiming superior performance to doctors have fuelled hype amongst the public and press for accelerated implementation.[7,8] 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"*.

The methodology and risk of bias of studies behind such headlines has 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. First, 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. And second, well conducted and transparently reported randomized clinical trials (RCTs) that evaluate usefulness in the real-world. Both 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 methodology and quality of reporting and transparency of deep learning studies that compare diagnostic algorithm performance to human clinicians with a view to suggesting 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) prior to search execution with any deviations from the protocol detailed in the Supplementary Appendix. This manuscript has been prepared according to the guidelines by the PRISMA group and a checklist is available with the Supplementary Appendix.[13]

Study identification and inclusion criteria

We performed a comprehensive search using free-text terms for various forms of the keywords 'deep learning' and 'clinician' to identify eligible studies. The exact search strategy is listed in Appendix 1. The following 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 list of relevant publications.

Publications were selected for review if they satisfied the following inclusion criteria: a peer reviewed scientific report of original research, English language, assessed a deep learning algorithm as applied to a clinical problem in medical imaging, compared algorithm performance to a contemporary human group not involved in establishing the ground truth (the true target disease status as verified by best clinical practice) and at least one human within the group was considered an expert. 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 via a convolutional neural network).[4] A clinical problem was defined as a situation in which a patient would ordinarily encounter a medical professional to improve or manage their health (this did not

include segmentation tasks e.g. delineating the borders of a tumour to calculate tumour volume). An expert was defined as an appropriately board-certified specialist/attending or equivalent.

Study selection and extraction of data

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

Adherence to reporting standards and risk of bias

For non-randomized studies, we assessed reporting quality of studies against a modified version of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) 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 including all sub-items) but some items were deemed less relevant to deep learning studies (e.g. points relating to predictor variables). We therefore used a modified list of 29 total points (see Appendix 2). The aim was to assess whether studies broadly conformed to reporting recommendations contained in TRIPOD, and not the detailed granularity required for a full assessment of adherence.[16] For the non-randomized studies, we assessed risk of bias by applying the Prediction model Risk Of Bias ASsessment Tool (PROBAST).[17] PROBAST contains 20 signalling questions from 4 domains (participants, predictors, outcomes and analysis) to allow assessment of the risk of bias in predictive modelling studies.[18] We did not assess applicability (as there was no specific therapeutic question for this systematic review) or predictor variables (as these are less relevant in deep learning studies on medical imaging) (see Appendix 2). For

randomized studies we assessed the broad level reporting against the Consolidated Standards of Reporting Trials (CONSORT) statement and risk of bias by applying the Cochrane Risk of Bias tool.[11,19]

Data synthesis

We intentionally planned not to conduct formal quantitative syntheses given the likely 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. This study was a systematic review of existing published research and therefore there was no active patient recruitment for data collection.

BOX of terms

Internal validation: Evaluation of model performance with data that used in the development process.

External validation: Evaluation of model performance with separate data not used in the development process.

Cross-validation: An internal validation approach in which data is randomly split into *n* equally sized groups. The model is developed in *n-1* of the *n* groups, and its performance evaluated in the remaining group with the whole process repeated *n* times and model performance taken as the average over the *n* iterations.

Bootstrapping: An internal validation approach which is similar to cross-validation but relies on ransom sampling with replacement.

Split-sample: An internal validation approach in which the available development data set is divided into 2 data sets – one to develop the model and the other to validate the model. The division can be random or non-random.

RESULTS

Study selection

A total of 8,302 records were retrieved by the electronic search last updated on 17 June 2019 (7,334 study records and 968 trial registrations, see Figure 1). Of the 7,334 study records, 140 full texts were assessed of which 59 were excluded. This left 81 non-randomized studies for analysis. Of the 968 trial registrations, 96 were assessed in full of which 86 were excluded leaving 10 trial registrations relating to deep learning.

Randomized clinical trials

The 10 trial registrations are summarised in Table 1. Eight related to gastroenterology and one each to ophthalmology and radiology. Eight were from China with one each from the USA and 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 attending ophthalmology clinics in China undergoing 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 of cataract diagnosis and treatment recommendation were 87% and 71% respectively, for the AI, which were significantly lower than 99% and 97% respectively, for senior consultants (p<0.001 for both) and also lower than the same AI when tested in a non-RCT setting (98% and 93% respectively). The mean time for receiving a diagnosis from the AI was faster than consultants (2.8 min vs. 8.5 min, p<0.001) and the authors suggested this might explain why patients were more satisfied with the AI (mean satisfaction score 3.47 vs. 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 1,058 patients undergoing colonoscopy with or without assistance of a real-time automatic polyp detection system that provided simultaneous visual and sound alerts upon encountering a polyp.[21] The authors found that the detection system resulted in a significant increase in the adenoma detection rate (29% vs. 20%, p<0.001), as well as an increase in the number of hyperplastic polyps found (114 vs. 52, *p*<0.001). Risk of bias was low in all domains except for blinding of participants, personnel and outcome assessors. Of note, one of the other trial registrations belongs to the same author group who are performing a double-blind RCT with sham AI to overcome the aforementioned blinding issue. 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-randomized studies – general characteristics

Nine of 81 non-randomized studies were prospective (11%) but only six of these nine were tested in a real-world clinical environment. The USA and Asia accounted for 82% of studies with the top four countries as follows: USA 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 of 81 (22%) studies compared how long a task took in both AI and human arms in addition to accuracy/performance metrics. Funding was predominantly academic (47/81, 58%) as opposed to commercial (9/81, 11%) or mixed (1/81, 1%). 12/81 studies stated they had no funding and another 12 did not report on funding. A detailed table with information on the 81 studies is included as a supplementary electronic file available online.

Seventy-seven of 81 studies made a specific comment in the abstract on 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 into the abstract that further prospective trials were required (absent in all 23 studies claiming superior performance to humans). 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 31/81 (38%) of studies. Seven of 81 (9%) of studies claimed in the discussion that the algorithm could now be used in clinical practice despite only two having been tested prospectively in a real-world setting. Concerning reproducibility, data were public and available in only 4/81 studies (5%). Code (for both pre-processing of data and modelling was available in only 6/81 studies (7%). Both raw labelled data and code were available in only one study.[22]

Non-randomized studies – methodology and risk of bias

Most studies both developed and validated a model (63/81, 78%) compared to development only with validation through resampling (9/81, 11%) or validation only (9/81, 11%). Where validation occurred in a separate dataset, this dataset was from a different geographical region in 19/35 studies (54%), from a different time period in 11/35 (31%) and a combination of both in 5/35 (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). Sample size calculations were reported in 14/81 studies (17%). Dataset sizes, where reported, were as follows: training (median 2,678, inter-quartile range (IQR) 704 to 21,362), validation (median 600, IQR 200 to 1,359) and test (median 337, IQR 144 to 891). The median event rate for development, validation and test sets was 42%, 44% and 44% respectively in cases where a binary outcome was assessed (n=62) as opposed to a multi-class classification (n=19). Forty-one of 81 studies used data augmentation (e.g. flipping and inverting images) to increase the dataset size.

The human comparator group was generally small (median 5, IQR 3 to 13, range 1 to 157) with a smaller group of experts within this (median 4, IQR 2 to 9, range 1 to 91). The number of participating non-experts varied from 0 to 94 (median 1, IQR 0 to 3). 36 of 81 studies used exclusively experts but in the 45 studies where non-experts were included 41 papers had some separate performance data available for exclusively the expert group. In the vast majority of studies, every human (expert or non-expert) rated the test dataset independently (blinded to all other clinical information except the image in 33/81 cases). The volume and granularity of the separate data for experts varied considerably between studies with some reporting individual performance metrics for each human (usually in supplementary appendices).

The overall risk of bias assessed using PROBAST led to 58/81 (72%) studies being classified as high risk (Figure 2) with the analysis domain being the most commonly rated 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).

Non-randomized studies – adherence to reporting standards

Adherence to reporting standards was poor (<50% adherence) for 12/29 TRIPOD items (see Figure 3). Overall, publications adhered to between 24% and 90% of the items of the TRIPOD statement with a median of 62% (IQR 45% to 69%). Eight TRIPOD items were reported in 90% or more of the 81 studies, and five items in less than 30% (Figure 3). A flow chart for the flow of patients/data through the study was only present in 25/81 studies (31%). Though not specifically requested in TRIPOD, we also looked for reporting of the hardware for developing or validating the algorithm which was reported in only 29/81 studies (36%)

and in the vast majority of cases (n=18) this related only to the graphics processing unit rather than full details (e.g. random access memory, central processing unit speed, configuration settings etc.).

DISCUSSION

We have conducted an appraisal of the methodology, adherence to reporting standards, risk of bias and claims of deep learning studies that compare diagnostic AI performance to clinicians. The rapidly advancing nature, and commercial drive of this field creates strong pressures to bring AI algorithms into clinical practice as quickly as possible. The potential consequences for patients of this implementation occurring 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]

There are five key findings from our review. First, we found very few relevant RCTs (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 does mean that claims about performance against clinicians should be tempered accordingly. However, given that deep learning only came into the mainstream in 2014, giving a lead-time of approximately five years for its testing within clinical environments, and that prospective studies may take a minimum of 1-2 years to conduct, it is reasonable to assume that many similar trials will be forthcoming over the next decade. We found only one randomized trial registered in the USA despite at least 16 deep learning algorithms approved for marketing by the Food and Drug Administration (FDA) in medical imaging covering a range of fields from radiology to ophthalmology and cardiology.[2,25]

Second, of the non-randomized studies, only nine were tested prospectively and just six performed testing in a real-world clinical environment. This makes comparisons of performance against clinicians difficult to evaluate given the artificial in silico context in which the clinician is being evaluated. In much the same way that surrogate endpoints do not always reflect clinical benefit,[26] a higher area under the curve may not necessarily lead to clinical benefit and may even have unintended adverse effects, such as an unacceptably

high false positive rate, that 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] The FDA has however recognised and acknowledged that their traditional paradigm of medical device regulation was not designed for adaptive artificial intelligence and machine learning technologies. Non-inferior AI (rather than superior) performance that allows for a lower burden on clinician workflow (i.e. being quicker with similar accuracy) may 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 an RCT (or at the very least prospective) setting. Even in an RCT 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?

Third, limited availability of datasets and code make assessing the reproducibility of deep learning research difficult to ascertain. Descriptions of the hardware used, where present, were also brief and this vagueness may 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 in the case of commercial concerns about intellectual property, there are strong arguments for ensuring that algorithms are non-proprietary and available for scrutiny.[32] This could be achieved by commercial companies collaborating with non-profit third parties for independent prospective validation.

Fourth, 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 and an appropriately large human sample for comparison is therefore 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 explicitly clear.

Fifth, descriptive phrases suggesting at least comparable (or better) diagnostic performance of an algorithm to a clinician were found in most abstracts, despite studies suffering from overt limitations in design, reporting, transparency and risk of bias. Caveats regarding the need for further prospective testing were rarely mentioned in the abstract (and not at all in the 23 studies claiming superior performance to a clinician). Accepting that abstracts are usually very 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 instead 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. It is clearly beyond the power of authors to control how the media and public interpret their findings but judicious and responsible use of language in studies and press releases, factoring in the strength and quality of the evidence, can help.[34] This issue is especially concerning given the findings from new research suggesting patients are more likely to consider a treatment beneficial when news stories are reported with spin, and that false news spreads much faster than true news online.[35,36]

The impetus for guiding best practice has gathered pace in the last year with a report proposing 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 likely 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 perhaps some degree of enforcement by journals in a similar vein to pre-registering randomized trials and reporting them

according to the CONSORT statement.[11,39] There is enthusiasm to speed up the process by which medical devices featuring 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 may help avoid hype.

Our findings must be considered in light of several limitations. First, although comprehensive, our search may nonetheless have missed some potentially includable studies. Second, the guidelines that we assessed non-randomized studies against (namely TRIPOD and PROBAST) were designed for conventional prediction modeling studies and so the adherence levels we found should be interpreted in this context. Third, we focused specifically on deep learning for medical imaging and so generalizing our findings to other types of AI such as conventional machine learning (for example, an artificial neural network based mortality prediction model using electronic health record data) may not be appropriate. Moreover, nomenclature in the field is sometimes used in non-standardized ways and thus some potentially eligible studies may have been presented with terminology that did not lead to them being captured with our search strategy. Fourth, risk of bias entails some subjective judgement and people with different prior experiences on AI performance may vary in their perceptions.

In conclusion, deep learning AI is an innovative and fast-moving field with 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 there are many arguably exaggerated claims regarding equivalence with (or superiority over) clinicians, which presents a potential risk for patient safety and population health at the societal level. Overpromising language leaves studies susceptible to being misinterpreted by the media and the public, and as a result the possible provision of inappropriate care that does not necessarily align with patients' best interests. Maximising

patient safety will be best served by ensuring that we develop a high quality and transparently reported evidence base moving forward.

AUTHOR CONTRIBUTIONS

MN & 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.

ACKNOWLEDGEMENTS

We thank Dina Radenkovic (DR) 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.

FUNDING AND SPONSORSHIP

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

CAL works 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. ACG 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 Labs, Myokardia and Voxel Cloud and the board of directors of Dexcoman and is an advisor to Guardant Health, Blue Cross Blue Shield Association, and Walgreens. MM is a co-founder of Cera, a technology-enabled homecare provider, Board Member of the NHS Innovation Accelerator, and Senior Advisor to Bain & Co.

TRANSPARENCY DECLARATION

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.

DATA SHARING

Raw data is available on request from the corresponding author.

EXCLUSIVE LICENSE

TABLE LEGENDS

Table 1. Randomized trial registrations of deep learning algorithms

FIGURE LEGENDS

Figure 1. Flow of study records (PRISMA diagram)

Figure 2. PROBAST risk of bias

Figure 3. Completeness of reporting of individual TRIPOD items

Table 1. Randomized 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-Jul-18	China	Gastroenterology	1000	AI-assisted colonoscopy	Standard colonoscopy	None	Polyp detection rate and adenoma detection rate	28-Feb-18

NCT03240848	Comparison of Artificial Intelligent Clinic and Normal Clinic	Completed, published	30-Jul-18	China	Ophthalmology	350	AI-assisted clinic	Normal clinic	Double (investigator and outcomes assessor)	Accuracy for congenital cataracts	25-May-18
NCT03706534	Breast Ultrasound Image Reviewed With Assistance of Deep Learning Algorithms	Recruiting	17-Oct-18	USA	Radiology	300	Computer-aided detection system	Manual ultrasound imaging review	Double (participant and investigator)	Concordance rate	31-Jul-19
NCT03840590	Adenoma Detection Rate Using AI System in China	Not yet recruiting	15-Feb-19	China	Gastroenterology	800	CSK AI system-assisted colonoscopy	Standard colonoscopy	None	Adenoma detection rate	01-Mar-20

NCT03842059	Computer-aided Detection for Colonoscopy	Not yet recruiting	15-Feb-19	Taiwan	Gastroenterology	1000	Computer-aided detection	Standard colonoscopy	Double (participant, care provider)	Adenoma detection rate	31-Dec-21
ChiCTR1800017675	The impact of a computer aided diagnosis system based on deep learning on incresing polyp detection rate during colonoscopy, a prospective double blind study	Not yet recruiting	21-Feb-19	China	Gastroenterology	1010	AI-assisted colonoscopy	Standard colonoscopy	Double	Polyp detection rate and adenoma detection rate	31-Jan-19

ChiCTR 1900021984	A multicenter randomized controlled study for evaluating the effectiveness of artificial intelligence in improving colonoscopy quality	Recruiting	19-Mar-19	China	Gastroenterology	1320	EndoAngel-assisted colonoscopy	Colonoscopy	Double (subjects and evaluators)	Polyp detection rate	31-Dec-20

NCT03908645	Development and Validation of a Deep Learning Algorithm for Bowel Preparation Quality Scoring	Not yet recruiting	09-Apr-19	China	Gastroenterology	100	AI-assisted scoring group	Conventional human scoring group	Single (outcome assessor)	Adequate bowel preparation	15-Apr-20
NCT03883035	Quality Measurement of Esophagogastroduodenoscopy Using Deep Learning Models	Recruiting	17-Apr-19	China	Gastroenterology	559	DCNN model-assisted EGD	Conventional EGD	Double (participant, care provider)	Detection of upper GI lesions	20-May-20

ChiCTR 1900023282	Prospective clinical study for artificial intelligence platform for lymph node pathology detection of gastric cancer	Not yet recruiting	20-May-19	China	Gastroenterology	60	Pathological Diagnosis of Artificial Intelligence	Traditional pathological diagnosis	Not stated	Clinical prognosis	31-Aug-21

REFERENCES

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/. Accessed September 22, 2019.

2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56.

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. Accessed September 22, 2019.

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

5. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019;25:24-9.

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

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/. Accessed September 22, 2019.

8. Price E. AI Is Better at Diagnosing Skin Cancer Than Your Doctor, Study Finds. 2018. http://fortune.com/2018/05/30/ai-skin-cancer-diagnosis/. Accessed September 22, 2019.

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.

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

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

12. Wallace E, Smith SM, Perera-Salazar R, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC Med Inform Decis Mak.* 2011;11:62.

13. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339: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.

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.

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.

17. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170:51-8.

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-W33.

19. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343: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.

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;68:1813-19.

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(1):e0191493.

23. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383:267-76.

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.

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/. Accessed September 22, 2019.

26. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;125:605-13.

27. FDA. Section 510(k) premarket notification of intent to market the device. 2018. https://www.accessdata.fda.gov/cdrh_docs/pdf18/K180647.pdf. Accessed September 22, 2019.

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. Accessed September 22, 2019.

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.

30. Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. *JAMA* 2014;312:1024-32.

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.

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.

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.

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.

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. *BMC Med* 2019;17:105.

36. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science* 2018;359:1146-51.

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. Accessed September 22, 2019.

38. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393: 1577-9.

39. De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;351:1250-1.

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(2):208-10.

41. Ratner M. FDA backs clinician-free AI imaging diagnostic tools. *Nature Biotech.* 2018;36:673.

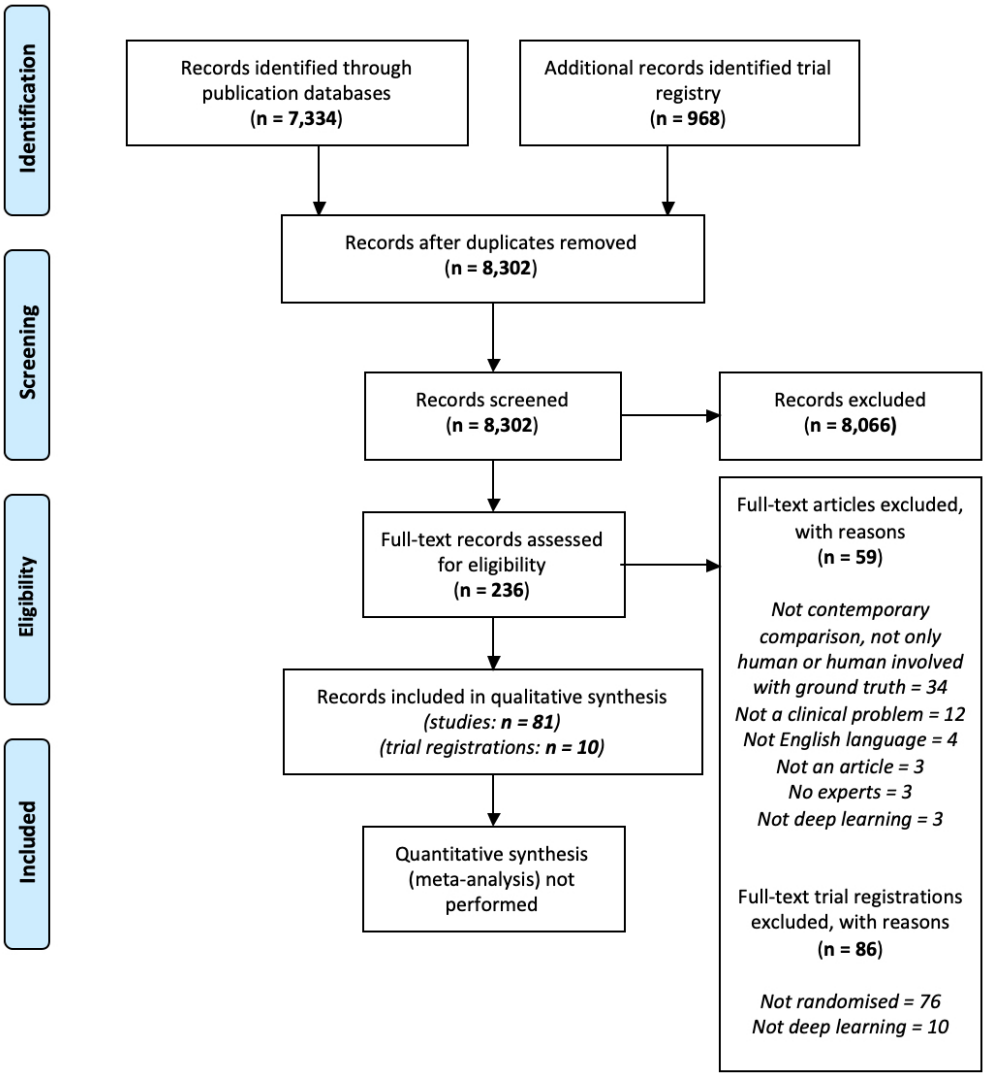

Figure 1. Flow of study records (PRISMA diagram)

182x197mm (144 x 144 DPI)

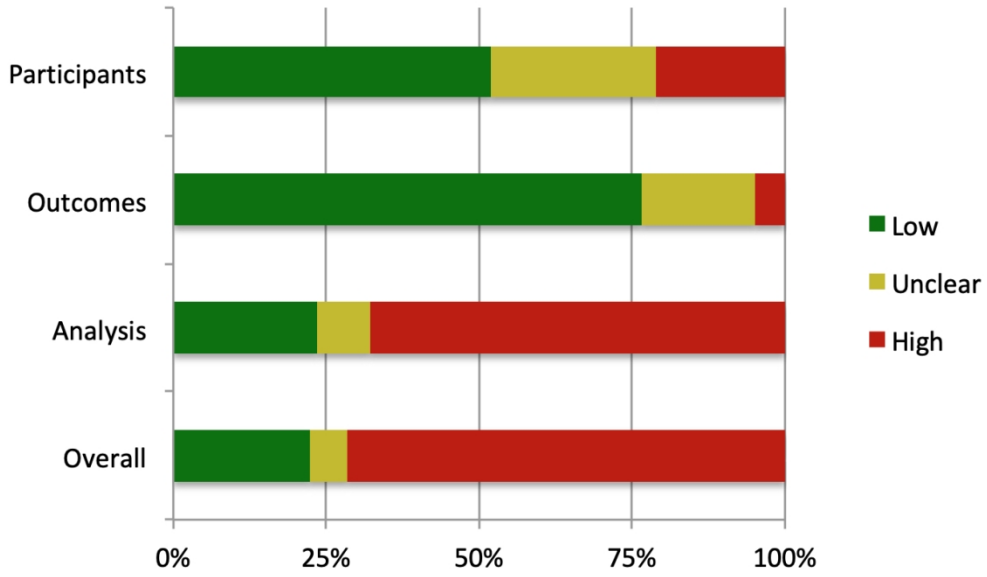

Figure 2. PROBAST risk of bias

245x151mm (144 x 144 DPI)

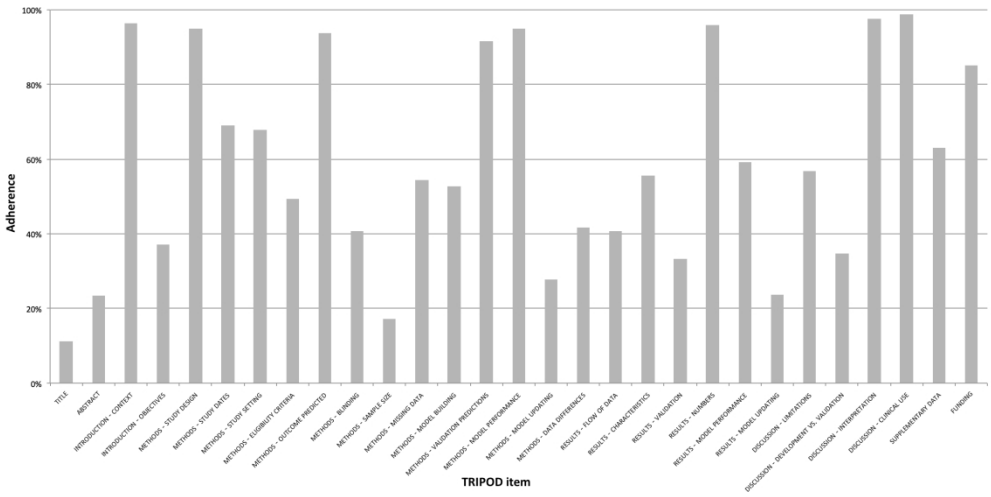

Figure 3. Completeness of reporting of individual TRIPOD items

394x196mm (144 x 144 DPI)

SUPPLEMENTARY APPENDICES:

Artificial intelligence vs. clinicians – a systematic review of the design, reporting standards, and claims of deep learning studies in medical imaging

Myura Nagendran[1]

Yang Chen[2]

Christopher A Lovejoy[3]

Anthony C Gordon[1,4]

Matthieu Komorowski[1,5]

Hugh Harvey[6]

Eric J Topol[7]

John P A Ioannidis[8]

Gary S Collins[9,10]

Mahiben Maruthappu[3]

1. ~~Section~~ 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. ~~Royal College of Radiologists~~Hardian Health, London, UK.
7. Scripps Research Translational Institute, La Jolla, California, USA.
8. Departments of Medicine, of Health Research and Policy, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, 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.

Version 4.1~~0.2~~

APPENDIX 1 – Search strategy

Medline and Embase (OvidSP)

1) artificial intelligence.ti OR AI.ti OR (neural network*).ti
2) machine learning.ti AND deep.ti,ab
3) ensemble.ti,ab AND deep.ti,ab
4) (deep learning OR deep-learning OR reinforcement learning OR reinforcement-learning OR deep neural network* OR deep belief network* OR convolutional neural network* OR recurrent neural network* OR feedforward neural network*).ti,ab
5) (Boltzmann machine* OR long short-term memory OR gated recurrent unit OR rectified linear unit OR autoencoder OR backpropagation OR multilayer perceptron OR convnet OR convolutional learning).ti,ab
6) 1 OR 2 OR 3 OR 4 OR 5
7) (board certified OR board-certified OR expert* OR expertise OR surgeon* OR clinician* OR physician* OR doctor* OR nurse* OR human* OR person* OR resident* OR attending* OR specialist* OR practitioner*).ti,ab
8) (anaesthesiologist* OR anaesthetist* OR cardiologist* OR dermatologist* OR endocrinologist* OR gastroenterologist* OR geriatrician* OR gynaecologist* OR haematologist* OR histopathologist* OR immunologist* OR intensivist* OR microbiologist* OR nephrologist* OR neurologist* OR neuroradiologist* OR obstetrician* OR oncologist* OR ophthalmologist* OR orthopaedic* OR otolaryngologist* OR paediatrician* OR pathologist* OR psychiatrist* OR pulmonologist* OR radiologist* OR rheumatologist* OR urologist*).ti,ab
9) (dietitian* OR echocardiographer* OR midwife* OR neurophysiologist* OR optometrist* OR paramedic* OR pharmacist* OR photographer* OR physiologist* OR physiotherapist* OR podiatrician* OR psychologist* OR radiographer* OR sonographer* OR therapist* OR ultrasonographer*).ti,ab
10) (inter observer OR inter-observer OR routine OR trial OR clinic).ti,ab
11) 7 OR 8 OR 9 OR 10
12) 6 AND 11
13) LIMIT 12 to yr=2010-2017
14) DEDUPLICATE 13
15) LIMIT 12 to yr=2018-2019

16) DEDUPLICATE 15

17) 14 OR 16

CENTRAL (Cochrane Central Register of Controlled Trials) (Wiley)

#1. (artificial intelligence):ti OR (AI):ti OR (neural network*):ti

#2. (machine learning):ti OR (ensemble):ti,ab,kw

#3. (deep):ti,ab,kw

#4. ((#2 AND #3)

#5. (deep learning OR deep-learning OR reinforcement learning OR reinforcement-learning OR deep neural network* OR deep belief network* OR convolutional neural network* OR recurrent neural network* OR feedforward neural network* OR Boltzmann machine* OR long short-term memory OR gated recurrent unit OR rectified linear unit OR autoencoder OR backpropagation OR multilayer perceptron OR convnet OR convolutional learning):ti,ab,kw

#6. #1 OR #4 OR #5

#7. (board certified OR board-certified OR expert* OR expertise OR surgeon* OR clinician* OR physician* OR doctor* OR nurse* OR human* OR person* OR resident* OR attending* OR specialist* OR practitioner* OR anaesthesiologist* OR anaesthetist* OR cardiologist* OR dermatologist* OR endocrinologist* OR gastroenterologist* OR geriatrician* OR gynaecologist* OR haematologist* OR histopathologist* OR immunologist* OR intensivist* OR microbiologist* OR nephrologist* OR neurologist* OR neuroradiologist* OR obstetrician* OR oncologist* OR ophthalmologist* OR orthopaedic* OR otolaryngologist* OR paediatrician* OR pathologist* OR psychiatrist* OR pulmonologist* OR radiologist* OR rheumatologist* OR urologist OR dietitian* OR echocardiographer* OR midwife* OR neurophysiologist* OR optometrist* OR paramedic* OR pharmacist* OR photographer* OR physiologist* OR physiotherapist* OR podiatrician* OR psychologist* OR radiographer* OR sonographer* OR therapist* OR ultrasonographer OR inter observer OR inter-observer OR routine OR trial OR clinic):ti,ab,kw

#8. #6 AND #7

#9. #8 with Publication Year from 2010 to 2019, in Trials

WHO ICTRP (available at http://apps.who.int/trialsearch/)

1) artificial intelligence OR machine learning OR deep learning OR algorithm OR neural network OR convolutional (search in Title, recruiting status: All)

APPENDIX 2 – TRIPOD and PROBAST altered or excluded items

TRIPOD items

TRIPOD item		Alteration for this study
1	*"Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted"*	Had to report whether study was development/validation/both, outcome of interest and mention deep learning or appropriate synonym in title
2	*"Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions"*	Predictors not applicable
3a	"*Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models"*	Deep learning model instead of multivariable prediction model
5c	*"Give details of treatments received, if relevant"*	Not assessed as unlikely to be applicable to deep learning studies
6b	"*Report any actions to blind assessment of the outcome to be predicted"*	We assessed whether any reporting on whether humans in test group blinded to other clinical data
7a	*"Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured"*	Not assessed as predictors not applicable
7b	*"Report any actions to blind assessment of predictors for the outcome and other predictors"*	Not assessed as predictors not applicable
9	"*Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method"*	Imputation not likely to be used in deep learning studies
10a	*"Describe how predictors were handled in the analyses"*	Not assessed as predictors not applicable
10b	"*Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation"*	Predictors not applicable
11	*"Provide details on how risk groups were created, if done"*	Not assessed as unlikely to be applicable to deep learning studies
12	*"For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors"*	Predictors not applicable
13a	"*Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful"*	Flow diagram/text/table can be at the level of analysis unit (e.g. flow of images rather than patients)
13b	*"Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome"*	Predictors not applicable

13c	*"For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome)"*	Predictors not applicable
14b	*"If done, report the unadjusted association between each candidate predictor and outcome"*	Not assessed as predictors not applicable
15a	*"Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point)"*	Not assessed as predictors not applicable
15b	*"Explain how to use the prediction model"*	Not assessed as predictors not applicable
18	*"Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data)"*	Predictors not applicable

PROBAST items

- Domain 1 – Participants
  - Risk of bias using all signalling questions
- Domain 2 – Predictors
  - Not assessed as not relevant for deep learning studies
- Domain 3 – Outcome
  - Risk of bias excluding signalling questions 3.3 and 3.5 as predictors not relevant for deep learning studies
- Domain 4 – Analysis
  - Risk of bias excluding signalling questions 4.2, 4.5 and 4.9 as predictors not relevant for deep learning studies
- Applicability sub-domains not assessed as no therapeutic question in this review

APPENDIX 3 – Protocol deviations

We described in our protocol in a section titled 'Protocol amendments' that: "*Given the rapidly developing landscape of this field, we anticipate that once the eligible studies are identified, there may be significant heterogeneity that means collection of some proposed data items are not feasible or extraction of alternative data items are preferable. Where any such changes occur, we will provide a clear rationale*."

Below we detail every deviation from the initially registered protocol along with rationale and possible limitations that may have arisen from these decisions. The protocol is available online at: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=123605

Protocol item	**Deviation**	**Rationale and potential limitations**
Methodological Expectations of Cochrane Intervention Reviews (MECIR standards)	Manuscript reported according to PRISMA guidelines only	This was felt to be sufficient given the lack of formal meta-analysis. Additional information is being made available in this supplement.
Eligibility criteria – studies	Letters included (but not letters to the editor)	The letter format of certain journals has a detail level on par with the full peer-reviewed reports of other journals. Limitation: the two letter studies in our sample should probably be given more slack regarding their adherence to TRIPOD.
Eligibility criteria – studies	Trial registrations had to be randomized to be included	The original protocol described observational or randomized trial registrations as being eligible (an error on our part and not our original intention). This was changed to only randomized trials as this was our main focus of interest in searching the trial registries.
Eligibility criteria – participants	Clinicians in the study had to be separate from any humans used to form the ground truth	This makes for a fairer comparison of performance as both AI and human are compared to a discrete and separate standard. Limitation: reduces the number of studies we could include but does mean that our sample is probably composed of more rigorous studies.
Eligibility criteria – participants	Experts did not have to be medically qualified to be an expert	In some fields, readers or graders (e.g. ophthalmology) may be specialist experts without being medically qualified. Limitation: might dilute the human performance standard.
Eligibility criteria – algorithm	Only studies in medical imaging with a convolutional neural network	The boundary between what constitutes a traditional machine learning algorithm with an artificial neural network and a true deep learning approach can be blurred. For clarity, we opted to focus on medical imaging studies with a convolutional neural network as we felt that this was where most deep learning studies

		with a human comparison would be published. Limitation: we may have missed deep learning studies in non-imaging areas such as deep reinforcement learning for treatment strategies. These are not usually evaluated against a separate human comparison however.
Outcomes	Data collected but not reported in paper (available on request)	We did not plan to perform meta-analyses. This makes sense given the highly heterogeneous nature of studies included but does mean that we are unable to make a claim about global performance between clinicians and AI. However, such a global metric would be so invalid as to probably be meaningless.
Search strategy	Authors of included studiesnot contacted to identify further studies	This was due to logistical constraints. Limitation: there is a very small chance we may have missed a few studies.
Search strategy	Search terms	The terms 'AI', 'neural network' and 'dermatologist' were missing from the original search strategy in the protocol. This was an error and corrected in the actually executed strategy listed in appendix 1. 2018 was also updated to 2019. Limitation: none.
Data collection	Commenting on efforts to prevent over-fitting not collected	This item was dropped in favour of items of more potential interest to clinicians. Limitation: if there is a major deficiency in reporting of over-fitting, we may not be able to comment on it.
Data collection	Proportion of missing data, imputation, imputation type, proportion of excluded data due to quality issues – not collected	Imputation was not relevant to deep learning studies. The other items on proportion of missing data were dropped in favour of items of more potential interest to clinicians. Limitation: we can make only anecdotal comments on the proportion of excluded data though this was also assessed in PROBAST question 4.3
Data collection	Whether or not humans were part of both comparator group and labeling was not collected	Studies with such humans would not be eligible for inclusion in the review. This item was listed in the protocol in error.
Data collection	Expertise level not collected	It was not felt to be reported clearly enough to make recording this data meaningful. Limitation: if there is a subtle difference between experts and renowned-experts, we would not be able to comment on it.
Data collection	Extra items not included in the protocol were collected	While reviewing studies for inclusion, it was felt that there were additional interesting items of data to collect of relevance to clinicians. An example incudes the coding of any comment on superiority of the AI over clinicians in the abstract. Limitation: our choice of items to collect in this regard was post-hoc and likely to be heavily influenced by what we were seeing. However, our aim was to highlight the most salient areas for improvement.
TRIPOD assessment	Not all 22 items used	See appendix 2

PROBAST assessment	Not all 20 signalling questions used	See appendix 2

APPENDIX 4 – Included studies

Randomized clinical trials

- 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. DOI 10.1136/gutjnl-2018-317500
- 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

Non-randomized studies

- Wang S, Wang R, Zhang S, et al. 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters <=3 cm using HRCT. Quantitative imaging in medicine and surgery 2018; 8(5): 491-9.
- Zhao W, Yang J, Sun Y, et al. 3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas. Cancer research 2018; 78(24): 6881-9.
- Matsuba S, Tabuchi H, Ohsugi H, et al. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. International ophthalmology 2018; (gsf, 7904294).
- Chen P-J, Lin M-C, Lai M-J, Lin J-C, Lu HH-S, Tseng VS. Accurate Classification of Diminutive Colorectal Polyps Using Computer-Aided Analysis. Gastroenterology 2018; 154(3): 568-75.
- Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. PloS one 2018; 13(3): e0193321.
- Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. npj Digital Medicine 2018; 1(1): 9.
- Poedjiastoeti W, Suebnukarn S. Application of Convolutional Neural Network in the Diagnosis of Jaw Tumors. Healthcare informatics research 2018; 24(3): 236-41.
- Zhu Y, Wang Q-C, Xu M-D, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointestinal endoscopy 2018; (0010505, fh8).
- Shichijo S, Nomura S, Aoyama K, et al. Application of Convolutional Neural Networks in the Diagnosis of Helicobacter pylori Infection Based on Endoscopic Images. EBioMedicine 2017; 25(101647039): 106-11.
- Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta orthopaedica 2019: 1-12.
- Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta orthopaedica 2017; 88(6): 581-6.

- Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature Biomedical Engineering 2017; 1: 0024.
- Hwang D-K, Hsu C-C, Chang K-J, et al. Artificial intelligence-based decision-making for age-related macular degeneration. Theranostics 2019; 9(1): 232-45.
- Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta orthopaedica 2018; 89(4): 468-73.
- Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. JAMA ophthalmology 2018; 136(7): 803-10.
- Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. JAMA ophthalmology 2017; 135(11): 1170-6.
- Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. European radiology 2019.
- Li F, Wang Z, Qu G, et al. Automatic differentiation of Glaucoma visual field from non-glaucoma visual filed using deep convolutional neural network. BMC medical imaging 2018; 18(1): 35.
- Wang L, Yang S, Yang S, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World journal of surgical oncology 2019; 17(1): 12.
- Kuo CC, Chang CM, Liu KT, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. npj Digital Medicine 2019; 2(1): 29.
- Byra M, Galperin M, Ojeda-Fournier H, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Medical physics 2018; (m82, 0425746).
- Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine 2019; 25(1): 65-9.
- Nakagawa K, Ishihara R, Aoyama K, et al. Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists. Gastrointestinal endoscopy 2019.
- Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. The British journal of radiology 2018; 91(1083): 20170576.
- Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. The Journal of investigative dermatology 2018; 138(7): 1529-38.
- De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine 2018; 24(9): 1342-50.
- Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. Computers in biology and medicine 2017; 82(doc, 1250250): 80-6.
- Ariji Y, Fukuda M, Kise Y, et al. Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of

artificial intelligence. Oral surgery, oral medicine, oral pathology and oral radiology 2018; (101576782).

- He Y, Guo J, Ding X, et al. Convolutional neural network to predict the local recurrence of giant cell tumor of bone after curettage based on pre-surgery magnetic resonance images. European radiology 2019.
- Brinker TJ, Hekler A, Enk AH, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. European journal of cancer (Oxford, England : 1990) 2019; 111: 148-54.
- Wu E, Hadjiiski LM, Samala RK, et al. Deep Learning Approach for Assessment of Bladder Cancer Treatment Response. Tomography (Ann Arbor, Mich) 2019; 5(1): 201-8.
- Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific reports 2018; 8(1): 3395.
- Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine 2018; 15(11): e1002686.
- Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. The Lancet Respiratory medicine 2018; 6(11): 837-45.
- Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. European radiology 2019; 29(7): 3338-47.
- Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: Detection of findings and presence of change. PloS one 2018; 13(10): e0204155.
- Kim Y, Lee KJ, Sunwoo L, et al. Deep Learning in Diagnosis of Maxillary Sinusitis Using Conventional Radiography. Investigative radiology 2019; 54(1): 7-15.
- Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. Investigative radiology 2017; 52(7): 434-40.
- Brinker TJ, Hekler A, Enk AH, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European journal of cancer (Oxford, England : 1990) 2019; 113: 47-54.
- Zucker EJ, Barnes ZA, Lungren MP, et al. Deep learning to automate Brasfield chest radiographic scoring for cystic fibrosis. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society 2019.
- Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. npj Digital Medicine 2019; 2(1): 25.
- Park A, Chute C, Rajpurkar P, et al. Deep Learning-Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. JAMA network open 2019; 2(6): e195600.
- Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images. Translational vision science & technology 2018; 7(6): 41.

- 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(1): e0191493.
- Nirschl JJ, Janowczyk A, Peyster EG, et al. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. PloS one 2018; 13(4): e0192726.
- Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS medicine 2018; 15(11): e1002699.
- Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. The British journal of dermatology 2018; (aw0, 0004041).
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639): 115-8.
- Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal radiology 2019; 48(2): 239-44.
- Rodriguez-Ruiz A, Krupinski E, Mordang J-J, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019; 290(2): 305-14.
- Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. npj Digital Medicine 2019; 2(1): 48.
- Ting DSW, Cheung CY-L, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 2017; 318(22): 2211-23.
- Choi KJ, Jang JK, Lee SS, et al. Development and Validation of a Deep Learning System for Staging Liver Fibrosis by Using Contrast Agent-enhanced CT Images in the Liver. Radiology 2018; 289(3): 688-97.
- Hwang EJ, Park S, Jin K-N, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA network open 2019; 2(3): e191095.
- Hwang EJ, Park S, Jin K-N, et al. Development and Validation of a Deep Learning-Based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2018; (a4j, 9203213).
- Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. Cancer Communications 2018; 38(1): 59.
- Nam JG, Park S, Hwang EJ, et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 2019; 290(1): 218-28.
- Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. The Lancet Oncology 2018; (100957246).

- Cha KH, Hadjiiski PhD LM, Cohan Md RH, et al. Diagnostic Accuracy of CT for Prediction of Bladder Cancer Treatment Response with and without Computerized Decision Support. Academic radiology 2018; (clv, 9440159).
- Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017; 318(22): 2199-210.
- Fujioka T, Kubota K, Mori M, et al. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. Japanese journal of radiology 2019; 37(6): 466-72.
- Choi JS, Han BK, Ko ES, et al. Effect of a Deep Learning Framework-Based Computer-Aided Diagnosis System on the Diagnostic Performance of Radiologists in Differentiating between Malignant and Benign Masses on Breast Ultrasonography. Korean journal of radiology 2019; 20(5): 749-58.
- Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nature Biomedical Engineering 2018; ((Lee, Yune, Mansouri, Kim, Tajmir, Guerrier, Ebert, Pomerantz, Romero, Kamalian, Gonzalez, Lev, Do) Department of Radiology, Massachusetts General Hospital, Boston, MA, United States).
- van Grinsven MJJP, van Ginneken B, Hoyng CB, Theelen T, Sanchez CI. Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images. IEEE transactions on medical imaging 2016; 35(5): 1273-84.
- Krause J, Gulshan V, Rahimy E, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology 2018; 125(8): 1264-72.
- Steiner DF, MacDonald R, Liu Y, et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. The American journal of surgical pathology 2018; 42(12): 1636-46.
- Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 2017; 35(c8s, 9713490): 303-12.
- Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of oncology : official journal of the European Society for Medical Oncology 2018; 29(8): 1836-42.
- Chee CG, Kim Y, Kang Y, et al. Performance of a Deep Learning Algorithm in Detecting Osteonecrosis of the Femoral Head on Digital Radiography: A Comparison With Assessments by Radiologists. AJR American journal of roentgenology 2019: 1-8.
- Gulshan V, Rajan RP, Widner K, et al. Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. JAMA ophthalmology 2019.
- Cha MJ, Chung MJ, Lee JH, Lee KS. Performance of Deep Learning Model in Detecting Operable Lung Cancer with Chest Radiographs. Journal of Thoracic Imaging 2019; ((Cha, Chung, Lee, Lee) Department of Radiology, Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea).

- Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine 2018; 1(1): 39.
- Ye H, Gao F, Yin Y, et al. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. European radiology 2019.
- Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PloS one 2017; 12(6): e0178992.
- Kise Y, Ikeda H, Fujii T, et al. Preliminary study on the application of deep learning system to diagnosis of Sjogren's syndrome on CT images. Dento maxillo facial radiology 2019: 20190019.
- Mori Y, Kudo S-E, Misawa M, et al. Real-Time Use of Artificial Intelligence in Identification of Diminutive Polyps During Colonoscopy: A Prospective Study. Annals of internal medicine 2018; 169(6): 357-66.
- Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. Journal of the American Academy of Dermatology 2018; 78(2): 270-7.e1.
- Zhang C, Sun X, Dang K, et al. Toward an Expert Level of Lung Cancer Detection and Classification Using a Deep Convolutional Neural Network. The oncologist 2019.
- Ciompi F, Chung K, van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Scientific reports 2017; 7(101563288): 46479.
- Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Use of Deep Learning for Detailed Severity Characterization and Estimation of 5-Year Risk Among Patients With Age-Related Macular Degeneration. JAMA ophthalmology 2018; 136(12): 1359-66.
- Sayres R, Taly A, Rahimy E, et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. Ophthalmology 2019; 126(4): 552-64.

APPENDIX 5 – Full results summary

Reporting quality and risk of bias – 2 studies (RCT)

- Wang et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019
    - Well reported on the whole
    - CONSORT checklist not included or referenced, however adherence to **30 of 37 points (81%)**

- Risk of bias assessed as per Cochrane risk of bias tool:
    - Random sequence generation LOW
    - Allocation concealment LOW
    - **Blinding of participants and personnel HIGH**
    - **Blinding of outcome assessors HIGH**
    - Incomplete outcome data LOW
    - Selective outcome reporting LOW
    - Other bias LOW

- The same group has another RCT in progress which is double blind with a sham AI to overcome the above issue

- Lin 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
    - Well reported on the whole
    - CONSORT checklist included, adherence to **31 of 37 points (84%)**

- Risk of bias assessed as per Cochrane risk of bias tool:
    - Random sequence generation LOW
    - Allocation concealment LOW
    - **Blinding of participants and personnel HIGH**
    - Blinding of outcome assessors LOW
    - Incomplete outcome data LOW
    - Selective outcome reporting LOW
    - Other bias LOW

Study characteristics – 81 studies (non-RCTs)

- **Prospective**: 9/81 (11%)
- **Prospective & real world testing**: 6/81 (7%)

- **Year**
  - 2016: 1 (1%)
  - 2017: 13 (16%)
  - 2018: 39 (48%)
  - 2019: 28 (35%)

- **Continent**
  - Asia: 42 (52%)
  - N. America 24 (30%)
  - Europe 15 (19%)

- **Country *(top 4)***
  - USA 24 (30%)
  - China 14 (17%)
  - South Korea 12 (15%)
  - Japan 9 (11%)

- **Specialty**
  - Radiology 36 (44%)
  - Ophthalmology 17 (21%)
  - Dermatology 9 (11%)
  - Gastroenterology 5 (6%)
  - Histopathology 5 (6%)
  - Orthopaedics 5 (6%)
  - Oncology 2 (2%)
  - Cardiology 1 (1%)
  - Nephrology 1 (1%)

- **ArXiv pre-print**
  - 5 prior to peer reviewed paper
  - 2 post peer-reviewed paper

- **Funding source**
  - Academic 47 (58%)
  - Commercial 9 (11%)
  - Mixed 1 (1%)
  - No funding 12 (15%)
  - Not reported 12 (15%)

- **Study type**
  - Development only 9/81 (11%)
  - Validation only 9/81 (11%)
  - Development & validation 63/81 (78%)
    - Validation in separate dataset 35/63 (61%)
      - Geographical 19/35 (54%)
      - Temporal (retrospective) 7/35 (20%)
      - Temporal (prospective) 4/35 (11%)
      - Geographical + temporal 5/35 (14%)

- **External dataset testing**
  - Foreign testing if external dataset used 20/32 (63%)

- **NICE digital health technology (DHT) type**
  - NICE recommends various standards of evidence for DHTs based on potential risk to user (full classification available at https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf
    - All 81 studies rated 3b

- **Internal validation method in studies not using a separate dataset**
  - *Some studies use >1 method*
    - Random split of dataset 18/37 (49%)
    - Non-random split of dataset 11/37 (30%)
    - Cross-validation 15/37 (41%)
    - Bootstrapping 6/37 (16%)

- **Ground truth**
    - Clinical ascertainment (c) 5/81 (6%)
    - Pathological (p) 25/81 (31%)
    - Human (h) 24/81 (30%)
    - Imaging report (i) 5/81 (6%)
    - Mixed (c / p / h / i) 22/81 (27%)

- **Training set**
  - Training set size
    - Available in 71 studies
    - Median 2,678 (IQR 704 to 21,362, range 56 to 1,665,151)

  - Training set events
    - *Event calculations only performed if binary outcome*
    - Available in 51 studies
    - Median 694 (IQR 200 to 3,500 range 23 to 131,731)
    - Proportion of events: median 42% (IQR 20% to 50%, range 2% to 81%)

- **Validation set**
  - Validation set size
    - Available in 37 studies
    - Median 600 (IQR 200 to 1,359, range 10 to 71,896)
  - Validation set events
    - *Event calculations only performed if binary outcome*
    - Available in 25 studies
    - Median 176 (IQR 85 to 300, range 5 to 28,637)
    - Proportion of events: median 44% (IQR 23% to 55%, range 2% to 79%)

- **Test set**
  - Test set size
    - Available in 74 studies
    - Median 337 (IQR 144 to 891, range 42 to 189,018)
  - Test set events
    - *Event calculations only performed if binary outcome*
    - Available in 54 studies
    - Median 139 (IQR 53 to 300, range 15 to 14,318)
    - Proportion of events: median 44% (IQR 23% to 58%, range 1% to 83%)

- **Human comparator group**
  - All humans: median 5 (IQR 3 to 13, range 1 to 157)
  - Experts: median 4 (IQR 2 to 9, range 1 to 91)

  - All human comparators are experts 36/81 (44%)
  - Some non-experts in comparator group 45/81 (56%)
    - Separate data available for expert group 41/45 (91%)

- **Availability of data**
  - Public, location provided and available 4/81 (5%)
  - Public, location provided but not all available 10/81 (12%)
  - Public, no location provided 18/81 (22%)
  - Unavailable or not reported 63/81 (78%)

- **Code availability**
  - Pre-processing of data 6/81 (7%)
  - Modelling 6/81 (7%)

- **Comment on algorithm vs. human clinician performance in abstract**
  - Algorithm superior 23/81 (28%) (23%)
  - Algorithm comparable or better 13/81 (16%) (16%)
  - Algorithm comparable 25/81 (31%) (33%)
  - Algorithm can help clinician perform better 14/81 (17%) (11%)
  - Algorithm not better 2/81 (2%) (4%)
  - No specific comment 4/81 (5%) (14%)

- **Abstract caveat of requirement for prospective +/- trials**
  - Reported in: 10/81 (12%)

- **Discussion caveat of requirement for further prospective work +/- randomised trials**
  - Reported in: 31/81 (38%)

- **Discussion states algorithm can be clinically used now**
  - Reported in: 7/81 (9%)

- **Comparison of algorithm vs. human clinician timing (*how long to perform task*)**
  - Reported in: 18/81 (22%)

- **Hardware that algorithm was tested on:**
  - Reported in: 29/81 (36%)
  - Where reported, was only graphical processing unit (GPU) in 18/29 (62%)

- **Data augmentation**
  - Used in: 41/81 (51%)

- **Study trial registry number**
  - Reported in: 7/81 (9%)
  - *However in 1 of these, the trial registry entry shows study as still recruiting and estimated study completion data is in the future. Trial registry was last updated AFTER the peer-reviewed paper was accepted for publication.*

- **Flow chart for participant / data flow** *(flow chart not in and of itself part of TRIPOD)*
  - Reported in: 25/81 (31%)

TRIPOD (reporting quality) – 81 studies (non-RCTs)

- **TRIPOD adherence:** studies adhered to median 62% of TRIPOD points (IQR 45 to 69, range 24 to 90)

- **TRIPOD study type**
  - 1a (development only) 0/81 (0%)
  - 1b (development and validation using resampling) 9/81 (11%)
  - 2a (random split-sample development and validation) 17/81 (21%)
  - 2b (non-random split-sample development and validation) 11/81 (14%)
  - 3 (development and validation using separate data) 35/81 (43%)
  - 4 (validation only) 9/81 (11%)

TRIPOD item	Development (D), validation (V) or both (DV)?	Adherence (%)	
1	DV	9/81	(11)
2	DV	19/81	(23)
3a	DV	78/81	(96)
3b	DV	30/81	(37)
4a	DV	77/81	(95)
4b	DV	56/81	(69)
5a	DV	55/81	(68)
5b	DV	40/81	(49)
6a	DV	76/81	(94)
6b	DV	33/81	(41)
8	DV	14/81	(17)
9	DV	44/81	(54)
10b	D	38/72	(53)
10c	V	66/72	(92)
10d	DV	77/81	(95)
10e	V	20/72	(28)
12	V	30/72	(42)
13a	DV	33/81	(41)
13b	DV	45/81	(56)
13c	V	24/72	(33)
14a	D	69/72	(96)
16	DV	48/81	(59)
17	V	17/72	(24)
18	DV	46/81	(57)
19a	V	25/72	(35)
19b	DV	79/81	(98)
20	DV	80/81	(99)
21	DV	51/81	(63)
22	DV	69/81	(85)

PROBAST (risk of bias) – 81 studies (non-RCTs)

- **Domain 1 – risk of bias in participants:**
  - Low: 42/81 (52%)
  - High: 17/81 (21%)
  - Unclear: 22/81 (27%)

- **Domain 2 – risk of bias in predictors:**
  - Not applicable

- **Domain 3 – risk of bias in outcome ascertainment:**
  - Low: 62/81 (77%)
  - High: 4/81 (5%)
  - Unclear: 15/81 (19%)

- **Domain 4 – risk of bias in analysis:**
  - Low: 19/81 (23%)
  - High: 55/81 (68%)
  - Unclear: 7/81 (9%)

- **OVERALL risk of bias:**
  - Low: 18/81 (22%)
  - High: 58/81 (72%)
  - Unclear: 5/81 (6%)

APPENDIX 6 – PRISMA checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	**1**
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	**3**
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	**4-5**
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	**4-5 and protocol**
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	**6**
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	**6 and protocol**
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	**6 and protocol**
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	**Appendix 1 and protocol**
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	**6**
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	**Protocol**
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	**Protocol**

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	**7**
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	**N/A**
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	**N/A**
Section/topic	#	Checklist item	**Reported on page #**
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	**N/A**
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	**N/A**
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	**8 and figure 1**
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	**Appendix 5**
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	**10**
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	**N/A**
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	**N/A**
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	**N/A**
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	**N/A**
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	**12-14**
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	**15**
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	**12, 15**
FUNDING			

Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	**16**

study_id	journal	pros_retro	pros_real	disc_caveat_p	use_clinically	funding
Abramoff	npj Digit. Med	1	1	0	1	c
Arbabshirani	npj Digit. Med	1	1	1	0	a
Arji	Oral Surg Oral	0	0	0	0	u
Becker 1	Invest Radiol	0	0	1	0	u
Becker 2	Br J Radiol	0	0	1	0	u
Bien	PLoS Med	0	0	0	0	n
Brown	JAMA Ophtha	0	0	0	0	a
Burlina 1	JAMA Ophtha	0	0	0	0	a
Burlina 2	JAMA Ophtha	0	0	0	0	a
Burlina 3	Comput Biol M	0	0	0	0	a
Bychov	Sci. rep.	0	0	0	0	a
Byra	Med Phys	0	0	0	0	a
Cha 1	J. Thorac. Ima	0	0	0	0	a
Cha 2	Acad Radiol	0	0	1	0	u
Chen	Gastroenterol	1	0	0	0	a
Choi 1	Radiology	0	0	0	0	a
Chung	Acta Orthop	0	0	1	0	a
Ciompi	Sci. rep.	0	0	0	0	a
De Fauw	Nat Med	0	0	1	0	u
Ehtesham Bej	JAMA	0	0	1	0	a
Esteva	Nature	0	0	1	0	a
Fujisawa	Br J Dermatol	0	0	1	0	n
Haenssle	Ann Oncol	0	0	1	0	n
Han 1	J Invest Derma	0	0	0	0	u
Han 2	PLoS ONE	0	0	0	0	c
Hannun	Nat Med	0	0	0	0	m
Hwang 1	Theranostics	0	0	0	1	a
Hwang 2	Clin Infect Dis	0	0	1	0	a
Kim	Invest Radiol	0	0	0	0	a
Kooi	Med Image An	0	0	0	0	a
Lee	Nat. Biomed. I	1	0	1	0	a
Li 1	Cancer Comm	1	0	0	1	a
Li 2	BMC med. ima	0	0	0	0	a
Li 3	Lancet Oncol	0	0	1	0	a
Lu 1	Transl. vis. sci	0	0	0	0	a
Marchetti	J Am Acad Der	0	0	1	0	a
Matsuba	Int Ophthalmo	0	0	0	0	u
Mori	Ann Intern Me	1	1	1	1	a
Nam	Radiology	0	0	1	0	a
Nirschl	PLoS ONE	0	0	0	1	a
Olczak	Acta Orthop	0	0	0	1	a
Poedjiastoeti	Healthc. infor	0	0	0	0	u
Rajpurkar	PLoS Med	0	0	1	0	n
Rodriguez-Rui	Radiology	0	0	1	0	c
Shichijo	EBioMedicine	0	0	0	0	n

Singh	PLoS ONE	0	0	0	0	n
Steiner	Am J Surg Patl	0	0	1	0	c
Ting	JAMA	0	0	1	0	a
Urakawa	Skeletal Radio	0	0	0	0	n
van Grinsven	IEEE Trans Me	0	0	0	0	a
Walsh	Lancet Respir	0	0	0	0	n
Wang 1	World J Surg (	0	0	1	0	n
Wang 2	Quant. imagin	0	0	0	0	a
Xue	PLoS ONE	0	0	0	0	c
Yu	PLoS ONE	0	0	0	0	a
Zhao	Cancer Res	0	0	0	0	a
Zhu	Gastrointest E	0	0	0	0	a
Brinker 1	European jour	0	0	0	0	n
Brinker 2	European jour	0	0	1	0	n
Chee	AJR. Americar	0	0	1	0	u
Choi 2	Korean journa	0	0	0	0	c
Ciritsis	European radi	0	0	0	0	n
Fujioka	Japanese jour	0	0	0	0	a
Gan	Acta orthopae	0	0	1	0	a
Gulshan	JAMA ophthal	1	1	1	0	c
Hamm	European radi	0	0	0	0	a
He	European radi	0	0	0	0	a
Hwang 3	JAMA networl	0	0	1	0	a
Kise	Dento maxillo	0	0	0	0	a
Kuo	npj Digital Me	0	0	0	1	a
Long	Nature Biome	1	1	0	0	a
Nagpal	npj Digital Me	0	0	0	0	u
Nakagawa	Gastrointestin	0	0	0	0	u
Park	JAMA networl	0	0	1	0	a
Raumviboons	npj Digital Me	1	1	1	0	u
Sayres	Ophthalmolog	0	0	0	0	c
Wu	Tomography (	0	0	0	0	a
Ye	European radi	0	0	1	0	a
Zhang	The oncologis	0	0	1	0	a
Zucker	Journal of cyst	0	0	1	0	a
Krause	Ophthalmolog	0	0	0	0	c

multi	train_set_size	train_set_eve	valid_set_size	valid_set_eve	test_set_size	test_set_even
.					819	198
.	37084	9939	9499		347	86
.	21362	10638				
.	786	178	395	90	321	53
.	445	62			192	20
.	1130	913	120	95	120	99
.	3575	684	915	176	100	15
1	59313		1348		6740	
.	133821	59420				
1	5664					
.	420	210			181	53
.	582	134	150	35	413	140
.	157	40				
.	19411	3500			500	200
.	2157	1476			284	188
1	7461				891	
.	1702	1238			189	138
1	1352		453		639	
1	14884		993		997	
.	270	110			129	49
1	127463				1942	
1	4867				1142	
.					300	60
1	17125				480	
.	49567	6673			1358	780
1	91232				328	
1	28720		7180		3872	
.	60089	6768	450	150	450	150
1	8000		1000		340	
.	39872	634	4218	85	18182	271
.	704	525	200	100	396	179
.	19576	13313	2690	1771	5270	3618
.	3712	1856			300	150
.	312399	131731			20386	10730
.	19815	5879	2202	653	837	537
.	900	173			379	75
.	253	95			111	42
.					466	287
.	42092	8625	600	300	600	300
.	104	47			105	47
.	179520	100228	51292	28637	25646	14318
.	400	200			100	50
1	98637		6351		420	
.					240	100
.	32208	13849			11481	2067

.						724	421
.						70	46
.		76370	2290	71896	2106	189018	10700
.		2678	1408	334	185	334	180
.		3959	655	1328	224	2592	609
.		929	602	89	59	139	92
.		5007	2557			564	351
.		1075	474	270	119	200	88
.		337	161	83	40		
.		724	350				
	1	523				128	
.		632	196	158	49	203	68
.		12378	1888	1359	230	100	20
.		12378	1888	1359	230	100	20
	1	1346		148		398	
.	.	.	.	.		253	80
	1	643		275		144	
.		947	467	.	.	120	72
.		2040	1341	.	.	300	150
.	.	.	.	.		5762	1380
	1	434		.		60	
.		56	23	.	.	.	.
.		87695	34074	1150	750	2104	1318
.		400	200	.	.	100	50
.		3609	.	401	.	495	.
.		886	410	.	.	110	54
	1	1226		.	.	331	
.		14338	2192	.	.	155	24
.		611	223	92	46	115	59
.	.	.	.	.		25323	3069
.	.	.	.	.		1796	213
.		77	58	10	5	42	30
.		2255	1461	282	181	299	194
.		2285	694	757	331	50	.
	1	1672		186		200	
	1	1665151		3737		1958	

ext_test	ext_test_forei	humans	experts	exp_which	prim_ai_meas	prim_ai_resul
1	0	.	.	.		
0	.	.	.	.	median time t	19 mins
0	.	2	2	.	AUC	0.8
1	1	3	2	1	AUC	0.81
0	.	3	1	1	AUC	0.84
1	1	9	9	.	accuracy	0.85
1	.	8	8	.	accuracy	0.91
0	.	1	1	.	linearly weigh	0.773
0	.	1	1	.	accuracy	0.916
0	.	1	1	.	weighted kapı	0.696
0	.	3	3	.	AUC	0.67
1	1	4	4	.	AUC	0.936
0	.	12	8	1	AUC	0.8
0	.	6	6	.	froc AUC	0.899
0	.	6	2	1	accuracy	0.901
1	0	4	3	1	obuchowski ir	0.94
0	.	58	30	1	accuracy	0.96
1	1	3	2	1	cohen-kappa	0.67
0	0	8	4	1	AUC	0.99
0	.	11	10	1	AUC	0.994
1	1	21	21	.	accuracy	0.72
0	.	22	13	1	accuracy	0.934
1	1	58	30	1	AUC	0.86
1	1	16	16	.	AUC	0.83 to 0.97
1	0	99	34	1	Youden index	0.676
1	0	6	6	.	Average F1 sc	0.837
1	1	4	2	1	accuracy	0.916
1	1	15	10	1	AUC	0.993
1	0	5	.	1	AUC	0.93, 0.88
0	.	3	2	1	AUC	0.852
0	.	5	2	1	AUC	0.961
0	.	14	3	1	accuracy	0.887
0	.	9	6	1	accuracy	0.876
1	0	6	6	.	accuracy	0.861
0	.	2	2	.	accuracy	0.952
0	.	8	8	.	AUC	0.86
0	.	6	.	0	AUC	0.9976
1	0	4	2	1	NPV	0.964
1	1	18	9	1	AUC	95.25
0	.	2	2	.	specificity	0.94
0	.	2	2	.	accuracy	0.83
0	.	5	5	.	accuracy	0.83
0	.	9	6	1	AUC	0.831
1	0	14	14	.	AUC	0.89
0	.	23	6	1	accuracy	0.877

1	1	4	4	.		AUC	0.843 to 0.93
1	1	6	6	.		sensitivity	0.91
1	1	3	1		0	AUC	0.936
0	.	5	2		0	accuracy	0.955
1	1	2	2	.		AUC	0.894 to 0.97
1	1	112	91		0	AUC	0.85
0	.	.	.			AUC	0.902
0	.	3	2		1	AUC	0.892
0	.	3	2		1	diagnosis agreement rate	
0	.	4	2		1	accuracy	0.8187
0	.	4	2		1	AUC	0.788
0	.	17	8		1	AUC	0.94
1	1	145	54		1	AUC	
0	.	157	63		1	AUC	
1	0	2	1		1	specificity	0.972
0	.	4	2		1	accuracy	0.921
1	1	2	2	.		accuracy	0.93
0	.	3	2		1	AUC	0.913
0	.	6	3		1	accuracy	0.93
1	1	4	2		1	specificity	0.952
0	.	2	2	.		accuracy	0.92
0	.	4	4	.		AUC	0.688
1	1	15	10		1	AUC	0.983
0	.	6	3		1	accuracy	0.96
0	.	4	4	.		accuracy	0.856
1	0	3	2		1	accuracy	
0	.	29	29	.		accuracy	0.7
0	.	16	16	.		accuracy	0.91
0	.	8	7		1	accuracy	0.809
1	1	13	13	.		specificity	0.934
0	.	10	9		1	specificity	0.947
0	.	2	2	.		AUC	0.73
0	.	4	1		1	AUC	1
1	0	25	25	.		accuracy	0.92
0	.	4	4	.			
0	.	3	3	.		quadratic wei	0.84

prim_exp_me	prim_exp_res	sec_ai_measu	sec_ai_result	sec_exp_mea	sec_exp_result
median time t	512 mins				
AUC	0.83	sensitivity	75.4	sensitivity	77.5
AUC	0.77 to 0.87	sensitivity	73.7	sensitivity	60 to 80
AUC	0.89	sensitivity	0.842	sensitivity	0.842
accuracy	0.894	sensitivity	0.879	sensitivity	0.905
accuracy	0.82	sensitivity	0.93		
linearly weigh	0.753	accuracy	0.757	accuracy	0.738
accuracy	0.902	sensitivity	0.884	sensitivity	0.864
weighted kapp	0.658	accuracy	0.794	accuracy	0.758
AUC	0.58				
AUC	0.849	sensitivity	0.848	sensitivity	0.992
AUC	0.74				
froc AUC	0.819	sensitivity	92	sensitivity	85
accuracy	0.8875	sensitivity	0.963	sensitivity	0.976
obuchowski ir	0.79				
accuracy	0.93	sensitivity	0.99	sensitivity	0.945
cohen-kappa	0.7				
		error rate	0.055	error rate	0.0675
AUC	0.966			sensitivity	0.628
accuracy	0.66	AUC	0.94-0.96		
accuracy	0.853	sensitivity	0.963		
mean ROC are	0.79	sensitivity	0.95	sensitivity	0.89
		sensitivity	77.7 to 98.6		
Youden index	0.484	sensitivity	0.96		
Average F1 sc	0.78				
AUC	0.959	sensitivity	0.952	sensitivity	0.929
AUC	0.86, 0.80	sensitivity	76.9%, 56.3%	sensitivity	77.6%, 59.2%
AUC	0.911				
		sensitivity	0.924		
accuracy	0.805	sensitivity	0.902	sensitivity	0.895
accuracy	0.606	sensitivity	0.932	sensitivity	0.62
accuracy	0.752	sensitivity	0.845	sensitivity	0.905
accuracy	0.932	sensitivity	0.94	sensitivity	0.951
AUC	0.71	sensitivity	0.52	sensitivity	0.82
accuracy	0.819	sensitivity	1	sensitivity	0.714
NPV	0.918	sensitivity	0.942	sensitivity	0.859
AUC	0.88	sensitivity	0.82325	sensitivity	0.704
specificity	0.78	sensitivity	0.99	sensitivity	0.725
accuracy	0.82				
accuracy	0.829	sensitivity	0.818	sensitivity	0.811
AUC	0.888	sensitivity	0.754	sensitivity	0.55
AUC	0.87	sensitivity	0.86	sensitivity	0.83
accuracy	0.889	sensitivity	0.889	sensitivity	0.852

AUC	0.723 to 0.845				
sensitvity	0.83	time per revie	61s	time per revie	117s
		sensitivity	0.905	sensitivity	91.2
accuracy	0.922	sensitivity	0.939	sensitivity	0.883
2		sensitivity	0.873	sensitivity	0.885
AUC	0.78	sensitivity	0.789	sensitivity	0.684
AUC	0.859	sensitivity	0.905	sensitivity	0.938
AUC	0.837	sensitivity	0.885	sensitivity	0.718
diagnosis agreement rate		sensitivity	0.95	sensitivity	1
accuracy	0.8136	sensitivity	0.9257	sensitivity	0.9659
		accuracy	0.641	accuracy	0.559
accuracy	0.775	sensitivity	0.7647	sensitivity	0.9081
AUC	0.761	sensitivity		sensitivity	0.912
AUC	0.685	sensitivity		sensitivity	0.73
specificity	0.861	sensitivity	0.752	sensitivity	0.78
accuracy	0.811	sensitivity	0.85	sensitivity	0.876
accuracy	0.953	sensitivity	0.895	sensitivity	0.921
AUC	0.843	sensitivity	0.958	sensitivity	0.889
accuracy	0.94	sensitivity	0.9	sensitivity	0.93
specificity	0.987	sensitivity	0.921	sensitivity	0.788
accuracy	0.88	sensitivity	0.9	sensitivity	0.83
AUC	0.753	sensitivity	0.857	sensitivity	0.583
AUC	0.914	sensitivity	0.913	sensitivity	0.844
accuracy	0.983	sensitivity	1	sensitivity	0.993
accuracy	0.705	precision	0.913	precision	0.912
accuracy		missed detect	4	missed detect	9.5
accuracy	0.61				
accuracy	0.896	sensitivity	0.901	sensitivity	0.898
accuracy	0.893	sensitivity	0.949	sensitivity	0.831
specificity	0.959	sensitivity	0.862	sensitivity	0.611
specificity	0.966	sensitivity	0.915	sensitivity	0.794
AUC	0.77	sensitivity	0.417		
AUC	1	sensitivity	0.99	sensitivity	1
accuracy	0.796	sensitivity	0.96	sensitivity	0.813
quadratic wei	0.82	sensitivity	0.971	sensitivity	0.755

Variable
study_id
journal
pros_retro
pros_real
tripod
dv
year
country
continent
disease
spec
outcome
human_abstract
abstract_caveat_prospective_or_trials
disc_caveat_prospective_or_trials
use_clinically
funding
multi
train_set_size
train_set_events
valid_set_size
valid_set_events
test_set_size
test_set_events
ext_test
ext_test_foreign
humans
experts
exp_which
prim_ai_measure
prim_ai_result
prim_exp_measure
prim_exp_result
sec_ai_measure
sec_ai_result
sec_exp_measure
sec_exp_result

Full name / description

First author of study
Journal of publication
Whether prospective or not
If prospective, tested in real clinical environment?
TRIPOD classification of study type
TRIPOD classification of development, validation or both
Year of publication
Country (of first and senior author)
Continent (of first and senior author)
Main disease/condition
Main specialty
Main outcome
Description of AI performance relative to humans in study abstract
Whether or not caveat made in abstract regarding need for further prospective work and/or trials
Whether or not caveat made in discussion regarding need for further prospective work and/or trials
Whether authors have advised using the algorithm in clinical practice
Funding source: a, academic; c, commercial; m, mixed; n, no funding; u, unclear
Was outcome binary or a multi-class classification?
Size of training set
Number of outcome events in training set
Size of validation set
Number of outcome events in validation set
Size of test set
Number of outcome events in test set
Was external testing used?
Was external testing dataset from different country?
Number of humans in comparator group
Number of experts in comaprator group
Was data reported separately for experts and non-experts?
Primary outcome (efficacy) measure for AI
Primary outcome (efficacy) result for AI
Primary outcome (efficacy) measure for human expert(s)
Primary outcome (efficacy) result for human expert(s)
Secondary outcome (safety) measure for AI
Secondary outcome (safety) result for AI
Secondary outcome (safety) measure for human expert(s)
Secondary outcome (safety) result for human expert(s)