SUPPLEMENTARY APPENDICES:

Artificial intelligence vs. clinicians – a systematic review of the design, reporting standards, and claims of deep learning studies in diagnostic 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. 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, 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 5.0

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

<u>PROBAST items</u>

- 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	Our aim in this paper was the comparison of AI against clinicians. We felt that the fairest comparison between AI and clinicians would be one in which both groups were compared to an independently ascertained gold standard and therefore the clinicians couldn't be involved in the establishment of this gold 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 studies not 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	There is potential for overfitting of a predictive algorithm to the index dataset used for development. However, 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	Our definition of expertise was: "an appropriately board-certified specialist/attending or equivalent". However, within this there could be an 'expert' who had only just become board-certified versus another who

		was an internationally renowned leader with several decades of clinical experience. This latter degree of detail was not consistently reported and not extracted by us during data collection. 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%)
  - This refers to a dataset that is completely separate from the index dataset. In some cases this was domestic data and in some cases a foreign dataset was obtained.

- **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%)
      - This could be any form of split that was not explicitly randomised (e.g. chronological or geographical).
    - 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%)

- **Not all studies split neatly into training, validation and testing. Some were development and internal validation only, others were external validation only. This can be appreciated more readily from the newly added supplementary electronic table. Sample size for studies with cross-validation or bootstrapping was the original number. The same applied for studies where data augmentation was used to boost sample size.**

- **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%)
  - Modelling refers to the code used to construct the actual deep learning algorithm (as distinct from pre-processing code used to sort and label the data prior to modelling).

- **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%)
  - Detailed information on the different data extraction techniques was not recorded during the data collection stage

- **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**