A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Xiaoxuan Liu*, Livia Faes*, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Brunyseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledesam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, Alastair K Denniston

Summary

Background Deep learning offers considerable promise for medical diagnostics. We aimed to evaluate the diagnostic accuracy of deep learning algorithms versus health-care professionals in classifying diseases using medical imaging.

Methods In this systematic review and meta-analysis, we searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019. Studies comparing the diagnostic performance of deep learning models and health-care professionals based on medical imaging, for any disease, were included. We excluded studies that used medical waveform data graphics material or investigated the accuracy of image segmentation rather than disease classification. We extracted binary diagnostic accuracy data and constructed contingency tables to derive the outcomes of interest: sensitivity and specificity. Studies undertaking an out-of-sample external validation were included in a meta-analysis, using a unified hierarchical model. This study is registered with PROSPERO, CRD42018091176.

Findings Our search identified 31 587 studies, of which 82 (describing 147 patient cohorts) were included. 69 studies provided enough data to construct contingency tables, enabling calculation of test accuracy, with sensitivity ranging from 0.7% to 100.0% (mean 79.1%, SD 0.2) and specificity ranging from 38.9% to 100.0% (mean 88.3%, SD 0.1). An out-of-sample external validation was done in 25 studies, of which 14 made the comparison between deep learning models and health-care professionals in the same sample. Comparison of the performance between health-care professionals in these 14 studies, when restricting the analysis to the contingency table for each study reporting the highest accuracy, found a pooled sensitivity of 87.0% (95% CI 83.0–90.2) for deep learning models and 86.4% (79.9–91.0) for health-care professionals, and a pooled specificity of 92.5% (95% CI 85.1–96.4) for deep learning models and 90.5% (80.6–95.7) for health-care professionals.

Interpretation Our review found the diagnostic performance of deep learning models to be equivalent to that of health-care professionals. However, a major finding of the review is that few studies presented externally validated results or compared the performance of deep learning models and health-care professionals using the same sample. Additionally, poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy. New reporting standards that address specific challenges of deep learning could improve future studies, enabling greater confidence in the results of future evaluations of this promising technology.

Funding None.

Copyright © 2019 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.
Research in context

Evidence before this study

Deep learning is a form of artificial intelligence (AI) that offers considerable promise for improving the accuracy and speed of diagnosis through medical imaging. There is a strong public interest and market forces that are driving the rapid development of such diagnostic technologies. We searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019, that developed or validated a deep learning model for the diagnosis of any disease feature from medical imaging material and histopathology, with no language restrictions. We prespecified the cutoff of Jan 1, 2012, to reflect a recognised change in model performance with the development of deep learning approaches. We found that an increasing number of primary studies are reporting diagnostic accuracy of algorithms to be equivalent or superior when compared with humans; however, there are concerns around bias and generalisability. We found no other systematic reviews comparing performance of AI algorithms with health-care professionals for all diseases. We did find two disease-specific systematic reviews, but these mainly reported algorithm performance alone rather than comparing performance with health-care professionals.

Added value of this study

This review is the first to systematically compare the diagnostic accuracy of all deep learning models against health-care professionals using medical imaging published to date. Only a small number of studies make direct comparisons between deep learning models and health-care professionals, and an even smaller number validate these findings in an out-of-sample external validation. Our exploratory meta-analysis of the small selection of studies validating algorithm and health-care professional performance using out-of-sample external validations found the diagnostic performance of deep learning models to be equivalent to health-care professionals. When comparing performance validated on internal versus external validation, we found that, as expected, internal validation overestimates diagnostic accuracy for both health-care professionals and deep learning algorithms. This finding highlights the need for out-of-sample external validation in all predictive models.

Methods

Search strategy and selection criteria

In this systematic review and meta-analysis, we searched for studies that developed or validated a deep learning model for the diagnosis of any disease feature from medical imaging material and histopathology, and additionally compared the accuracy of diagnoses made by algorithms versus health-care professionals. We searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019, with no language restrictions. The full search strategy for each database is available in the appendix (p 2). The cutoff of Jan 1, 2012, was prespecified on the basis of a recognised step-change in machine learning performance with the development of deep learning approaches. In 2012, for the first time, a deep learning model called AlexNet, enabled by advances in parallel computing architectures, made an important breakthrough at the ImageNet Large-Scale Visual Recognition Challenge. The search was first performed on and up to May 31, 2018, and an updated search was performed on June 6, 2019. Manual searches of bibliographies, citations, and related articles (PubMed function) of included studies were undertaken to identify any additional relevant articles that might have been missed by the searches.
Eligibility assessment was done by two reviewers who screened titles and abstracts of the search results independently, with non-consensus being resolved by a third reviewer. We did not place any limits on the target population, the disease outcome of interest, or the intended context for using the model. For the study reference standard to classify absence or presence of disease, we accepted standard-of-care diagnosis, expert opinion or consensus, and histopathology or laboratory testing. We excluded studies that used medical waveform data graphics material (ie, electroencephalography, electrocardiography, visual field data) or investigated the accuracy of image segmentation rather than disease classification.

Letters, preprints, scientific reports, and narrative reviews were included. Studies based on animals or non-human samples or that presented duplicate data were excluded.

This systematic review was done following the recommendations of the PRISMA statement. Methods of analysis and inclusion criteria were specified in advance. The research question was formulated according to previously published recommendations for systematic reviews of prediction models (CHARMS checklist).

Data analysis
Two reviewers (XL, then one of LF, SKW, DJF, AK, AB, or TM) extracted data independently using a predefined data extraction sheet, cross-checked the data, and resolved disagreements by discussion or referral to a third reviewer (LMB or AKD). We contacted four authors for further information. One provided numerical data that had only been presented graphically in the published paper and one confirmed an error in their published contingency table. We did not formally assess the quality of the included studies.

Where possible, we extracted binary diagnostic accuracy data and constructed contingency tables at the reported thresholds. Contingency tables consisted of true-positive, false-positive, true-negative, and false-negative results, and were used to calculate sensitivity and specificity.

To estimate the accuracy of deep learning algorithms and health-care professionals, we did a meta-analysis of studies providing contingency tables from out-of-sample external validations (including geographical and temporally split data). If a study provided various contingency tables for the same or for different algorithms, we assumed these to be independent from each other. We accepted this assumption because we were interested in providing an overview of the results of various studies rather than providing precise point estimates. We used a unified hierarchical model that was developed for the meta-analysis of diagnostic accuracy studies and plotted summary receiver operating characteristic (ROC) curves for the accuracy of health-care professionals and deep learning algorithms. The hierarchical model involves statistical distributions at two different levels. At the lower level, it models the cell counts that form the contingency tables (true positive, true negative, false positive, and false negative) by using binomial distributions. This accounts for the within-study variability. At the higher level, it models the between-study variability (sometimes called heterogeneity) across studies. The hierarchical summary ROC figures provide estimates of average sensitivity and specificity across included studies with a 95% confidence region of the summary operating point and the 95% prediction region, which represents the confidence region for forecasts of sensitivity and specificity in a future study.

Owing to the broad nature of the review—ie, in considering any classification task using imaging for any disease—we were accepting of a large degree of between-study heterogeneity and thus it was not formally assessed.

![Figure 1: Study selection](image-url)
| Subspecialty                  | Participants                                                                 | Inclusion criteria                                                                                           | Exclusion criteria                                                                                   | Mean age (SD; range), years | Percentage of female participants | Number of participants represented by the training data |
|------------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|----------------------------|-----------------------------------|------------------------------------------------------|
| Abbasi-Sureshjani et al (2018)† | Ophthalmology                                                               | NR                                                                                                           | NR                                                                                                    | NR (NR; 40–76)             | 51%                               | NR                                                   |
| Adams et al (2019)†          | Trauma and orthopaedics                                                      | Emergency cases of surgically confirmed neck of femur fractures                                               | Other radiological pathology present (excluding osteoporosis or osteoarthritis); metal-wear in fractured or unfractured hip | NR                        | NR                                | NR                                                   |
| Ardila et al (2019)†          | Lung cancer                                                                  | Lung cancer screening patients                                                                                  | Unmatched scans to radiology reports; patients >1 year of follow-up                                  | NR                        | NR                                | 12 504                                               |
| Ariji et al (2019)†           | Oral cancer                                                                  | Patients with contrast-enhanced CT and dissection of cervical lymph nodes                                      | NR                                                                                                    | Median 63 (NR; 33–95)      | 47%                               | NR                                                   |
| Ayed et al (2015)†            | Breast cancer                                                                | NR                                                                                                           | NR                                                                                                    | NR (NR; 24–88)             | 100%                              | NR                                                   |
| Becker et al (2017)†          | Breast cancer                                                                | Mammograms with biopsy proven malignant lesions                                                              | Surgery before first mammogram; metastatic malignancy involving breasts; cancer >2 years on external mammogram; in non-malignant cases, patients with <2 years of follow-up | 57 (9; 32–85)             | 100%                              | 2038                                                 |
| Becker et al (2018)‡          | Breast cancer                                                                | Mammograms with biopsy proven malignant lesions                                                              | Normal breast ultrasound or benign lesions, except if prior breast-conserving surgery was done; no radiological follow-up >2 years or histopathology proof | NR                        | 100%                              | NR                                                   |
| Bien et al (2018)‡            | Trauma and orthopaedics                                                      | NR                                                                                                           | NR                                                                                                    | NR                        | NR                                | NR                                                   |
| Brinker et al (2019)†         | Dermatological cancer                                                        | NR                                                                                                           | NR                                                                                                    | NR                        | NR                                | NR                                                   |
| Brown et al (2018)†           | Ophthalmology                                                                | NR                                                                                                           | Stage 4–5 retinopathy of prematurity                                                                      | NR                        | NR                                | 898                                                  |
| Burlina et al (2017)†         | Ophthalmology                                                                | NR                                                                                                           | NR                                                                                                    | NR                        | NR                                | NR                                                   |
| Burlina et al (2018)†         | Ophthalmology                                                                | NR                                                                                                           | NR                                                                                                    | NR                        | NR                                | NR                                                   |
| Burlina et al (2018)†         | Ophthalmology                                                                | NR                                                                                                           | NR                                                                                                    | NR                        | NR                                | NR                                                   |
| Byra et al (2019)‡            | Breast cancer                                                                | Masses with images in at least two ultrasound views                                                          | Inconclusive pathology; artifacts or known cancers                                                    | NR                        | NR                                | NR                                                   |
| Cao et al (2019)†             | Urology                                                                      | Patients undergoing robotic assisted laparoscopic prostatectomy with pre-operative MRI scans                  | Patients with prior radiotherapy or hormonal therapy                                                   | NR                        | NR                                | NR                                                   |
| Chee et al (2019)†            | Trauma and orthopaedics                                                      | Patients aged ≥16 years with hip pain with osteonecrosis of the femoral head on MRI                            | >30 days between anteroposterior hip x-ray and hip MRI; history of hip operation with osseous abnormality in femoral head and neck; insufficient MRI and poor radiograph quality | 48 (15; NR)               | 48%                               | NR                                                   |
| Choi et al (2019)†            | Breast cancer                                                                | Patients aged ≥20 years with breast masses on ultrasound                                                     | Undiagnosed breast mass and low-quality images                                                        | Median 47 (NR; 42–54)     | NR                                | NR                                                   |
| Choi et al (2018)†            | Hepatology                                                                   | Training set: pathologically confirmed cases                                                                 | Training dataset: 44 (15; 18–83); Test dataset 1: 48 (14, NR); Test dataset 2: 56 (10, NR); Test dataset 3: 53 (15, NR) | Training dataset: 28% | Total test datasets: 43% | 7461                                                 |

(Table 1 continues on next page)
Subspecialty | Participants | Inclusion criteria | Exclusion criteria | Mean age (SD; range), years | Percentage of female participants | Number of participants represented by the training data
--- | --- | --- | --- | --- | --- | ---
(Continued from previous page)

Ciompi et al (2017) 39 | Respiratory disease | Baseline CT scans from the Multicentric Italian Lung Detection trial | Lesion diameter <4 mm | NR | NR | 943
Codella et al (2017) 40 | Dermatological cancer | NR | NR | NR | NR | NR
Coudray et al (2018) 41 | Lung cancer | NR | NR | NR | NR | NR
De Fazio et al (2018) 42 | Ophthalmology | All routine OCT images | Conditions with <10 cases | NR | Training dataset: 54% Test dataset: 55% | 7621
Ding et al (2019) 43 | Neurology, psychiatry | Patients participating in the Alzheimer’s Disease Neuroimaging Initiative clinical trial | Patients with no PET study ordered | Male: 76 (NR; 55–93) Female: 75 (NR; 55–96) | 47% | 899
Dunnmon et al (2019) 44 | Respiratory disease | Images which are not anteroposterior or posteroanterior views | NR | NR | 200,000
Ehteshami Bejnordi et al (2017) 45 | Breast cancer | Patients having breast cancer surgery | Isolated tumour cells in a sentinel lymph node | NR | 100% | NR
Esteva et al (2017) 46 | Dermatological cancer | NR | NR | NR | NR | NR
Fujioka et al (2019) 47 | Breast cancer | Breast ultrasound of benign or malignant masses confirmed by pathology; patients with minimum 2-year follow-up | Patients on hormonal therapy or chemotherapy; patients aged <20 years | Training dataset: 55 (13; NR) Test dataset: 57 (15; NR) | NR | 237
Fujisawa et al (2019) 48 | Dermatological cancer | NR | NR | NR | NR | 1842
Gómez-Valverde et al (2019) 49 | Ophthalmology | Aged 55–86 years in glaucoma detection campaign | Poor-quality images | NR | NR | NR
Grewal et al (2018) 50 | Trauma and orthopaedics | NR | NR | NR | NR | NR
Haenssle et al (2018) 51 | Dermatological cancer | NR | NR | NR | NR | NR
Hamn et al (2019) 52 | Liver cancer | Untreated liver lesions, or treated lesions that showed progression, or recurrence post 1 year local or regional therapy | Atypical imaging features; patients aged <18 years | 57 (14; NR) | 48% | 296
Han et al (2018) 53 | Dermatological cancer | All images from datasets | For the Asan dataset, postoperative images were excluded | Asan 1: 47 (23; NR) Asan 2: 41 (21; NR) Atlas: NR MED-NODE: NR Hallym: 68 (13; NR) Edinburgh: NR | Asan 1: 55% Asan 2: 57% Atlas: NR MED-NODE: NR Hallym: 52% Edinburgh: NR | NR
Han et al (2018) 54 | Dermatological cancer | For Inje, Hallym, and Seoul datasets: onychomycosis: positive potassium, oxygen, and hydrogen test or fungus culture result; or successful treatment with antifungal drugs; nail dystrophy: negative potassium, oxygen, and hydrogen test or culture result; unresponsiveness to antifungal medication; or responsiveness to a triamcinolone intralesional injection | Inadequate images and images of uncertain diagnosis | Asan 1: 41 (22; NR) Asan 2: 46 (20; NR) Inje 1: 48 (23; NR) Inje 2: 54 (20; NR) Hallym: 39 (15; NR) Seoul: 51 (20; NR) | Asan 1: 55% Asan 2: 59% Inje 1: 56% Inje 2: 48% Hallym: 47% Seoul: 54% | NR
Hwang et al (2018) 55 | Respiratory disease | Active pulmonary tuberculosis ≤1 month from treatment initiation | Non-parenchymal tuberculosis and non-tuberculosis chest x-rays | 51 (16; NR) | 82% | NR

(Table 1 continues on next page)
| Subspecialty                  | Participants | Inclusion criteria                                                                 | Exclusion criteria                                                                 | Mean age (SD; range), years | Percentage of female participants | Number of participants represented by the training data |
|------------------------------|--------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|----------------------------|-----------------------------------|-------------------------------------------------------|
| (Continued from previous page) |              |                                                                                      |                                                                                     |                            |                                   |                                                       |
| Hwang et al (2019)56         | Ophthalmology| Age-related macular degeneration cases presenting to the hospital                    | Low-resolution images or improper format                                             | NR                         | NR                                | NR                                                    |
| Hwang et al (2019)57         | Respiratory  | Cases of clinically or microbiologically confirmed pneumonia or clinically reported pneumothorax; cases of pulmonary tuberculosis (where a chest x-ray was completed within 2 weeks of treatment initiation) | Chest x-rays >3 lesions for lung cancer; pneumothorax chest x-rays with drainage catheter or subcutaneous emphysema | Training dataset: 51 (16, NR) normal images; 62 (15, NR) for abnormal images | Training dataset: 55% | Test dataset: 38% |
| Kemany et al (2018)55         | Ophthalmology, respiratory disease | OCT: routine OCTs from local databases for choroidal neovascularisation, DMO, drusen, and normal images Chest x-rays: retrospective cohort of 1–5 year olds | OCT: none Choroidal neovascularisation 1: 83 (NR; 58–97) DMO 2: 57 (NR; 20–90) Drusen: 82 (NR; 40–95) Normal: 60 (NR; 21–68) X-ray: NR | Choroidal neovascularisation 1: 46% DMO 2: 62% Drusen: 56% Normal: 41% X-ray: NR | Training dataset: 4686 Chest x-ray: 5856 |
| Kim et al (2012)59            | Breast cancer | Patients with solid mass on ultrasound                                                | Breast Imaging Reporting and Data System: 0, 1, and 6                                | 44 (NR, 22–70)             | NR                                | 70                                                    |
| Kim et al (2018)60            | Trauma and orthopaedics | Tuberculous or pyogenic spondylitis                                                   | Tuberculosis spondylitis: 59 (NR, 38–71) Pyogenic spondylitis: 64 (NR, 56–72) | Tuberculosis spondylitis: 49% Pyogenic spondylitis: 40% | NR                                |                                                       |
| Kim et al (2019)61            | Maxillofacial surgery | Age >16 years with suspected maxillary sinusitis with a Waters’ view plain film radiographs | History of sinus surgery, fracture, or certain tumours involving the maxillary sinus | Training dataset: 47 (20, NR) Test dataset: internal validation: 54 (21, NR); external validation: temporal 49 (20, NR); geographical: 52 (18, NR) | Training dataset: 54% Test dataset: internal validation: 56%; external validation: temporal 47%, geographical 54% | NR                                                    |
| Kise et al (2019)62           | Rheumatology  | Sjogren’s syndrome                                                                  | Sjogren’s syndrome: 67 (NR, NR) Control: 66 (NR, NR)                                | Sjogren’s syndrome: 4% Control: 9.7% | 40                                |                                                       |
| Ko et al (2019)63             | Thyroid cancer | Ultrasound and subsequent thyroidectomy, nodules 1–2 cm with correlating pathology results | NR                                                                                  | Training dataset: 48 (13, 12–79) Test dataset: 50 (12, NR) | Training dataset: 82% Test dataset: 85% | NR                                                    |
| Kumagai et al (2019)64        | Oesophageal cancer | NR                                                                                  | NR                                                                                  | NR                          | 240                               |                                                       |
| Lee et al (2019)65            | Trauma and orthopaedics | Training and test data: non-contrast head CT with or without acute ICH Prospective test data: non-contrast head CT in 4 months from the local hospital’s emergency department | History of brain surgery, skull fracture, intracranial tumour, intracranial device, cerebral infarct, or non-acute ICH | NR                          | NR                                |                                                       |
| Li C et al (2018)66           | Nasopharyngeal cancer | Nasopharyngeal endoscopic images for screening                                       | Blurred images or images with incomplete exposure                                    | Training dataset: 46 (13, NR) Test dataset: 46 (13, NR) Prospective test dataset: 48 (13, NR) | Training dataset: 30% Test dataset: 32% Prospective test dataset: 34% | 5557                                                   |

(Table 1 continues on next page)
| Subspecialty                | Participants                                                                 | Mean age (SD; range), years | Percentage of female participants | Number of participants represented by the training data |
|----------------------------|-------------------------------------------------------------------------------|----------------------------|----------------------------------|-------------------------------------------------------|
| Thyroid cancer             | Patients aged ≥18 years with thyroid cancer; patients with pathological examination and negative controls | Training dataset: median 44 (NR; 36-54) | Training dataset: 75% | 42,952                                                  |
| Breast cancer              | Solid mass on ultrasound                                                      | NR                         | 100%                             | NR                                                    |
| Trauma and orthopaedics    | Routine examinations done as part of the Childhood Cataract Program of the Chinese Ministry of Health, and search engine images matching the key words “congenital”, “infant”, “paediatric cataract”, and “normal-eye” | NR                         | NR                               | NR                                                    |
| Ophthalmology              | Image containing only one of the four abnormalities (serous macular detachment, cystoid macular oedema, macular hole, and epiretinal membrane) | Images with other abnormalities than the four included or co-existence of two abnormalities | Control: 28% Wet age-related macular degeneration: 26% | NR                                                    |
| Oesophageal cancer         | Patients with superficial oesophageal squamous cell carcinoma with pathologic proof of cancer invasion depth | Severe oesophagitis; oesophageal chemotherapy or radiation history; lesions adjacent to ulcer or ulcer scar | Control: 77 (5; NR) Wet age-related macular degeneration: 76 (82; NR) | NR                                                    |
| Oral and maxillofacial cancer | Training: malignant lung nodules chest x-rays proven by histopathology External validation: chest x-rays with referential normal CTs performed within 1 month | Nodules ≤5 mm on CT, chest x-rays showing ≥3 nodules, lung consolidation, or pleural effusion obscuring view | Female: 52 (NR) Male: 53 (NR) | NR                                                    |
| Respiratory disease        | Panoramic x-rays of ameloblastomas and keratocystic odontogenic tumours with known biopsy results | Pathologies precluding classification of target condition, or presence of other retinal vascular disease | 61 (11; NR) | 67%                                                   |

(Continued from previous page)
### Table 1: Subspecialty Participants

| Subspecialty                  | Participants                                                                 | Inclusion criteria                                                                 | Exclusion criteria                                                                 | Mean age (SD; range), years | Percentage of female participants | Number of participants represented by the training data |
|-------------------------------|------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-----------------------------|----------------------------------|--------------------------------------------------------|
| (Continued from previous page) |                                                                              |                                                                                     |                                                                                    |                             |                                  |                                                         |
| Schlegl et al (2018)22        | Ophthalmology                                                                | Random sample of age-related macular degeneration, DMO, and retinal vein occlusion cases | No clear consensus or poor image quality                                           | NR                          | NR                               | NR                                                     |
| Shibutani et al (2019)81      | Cardiology                                                                   | Myocardial perfusion SPECT within 45 days of coronary angiography                   | NR                                                                                 | 72 (9; 50–89)               | 19%                              | NR                                                     |
| Shichijo et al (2017)82       | Gastroenterology                                                             | A primary care referral for OGD for epigastric symptoms, abnormal pepsinogen levels, previous gastroduodenal disease, or screening for gastric cancer | Helicobacter pylori eradication; presence or history of gastric cancer, ulcers, or submucosal tumour; unclear images | Training dataset: 53 (13; NR) Test dataset: 50 (11; NR) | Training dataset: 55% Test dataset: 57% | 735                                                     |
| Singh et al (2018)83          | Respiratory disease                                                           | Randomly selected chest x-rays from the database                                    | Lateral radiographs; oblique views; patients with total pneumonectomy; patients with a metal prosthesis | NR                          | NR                               | NR                                                     |
| Song et al (2019)84           | Thyroid cancer                                                                | Patients aged >18 years with total or nearly total thyroidectomy or lobectomy, with complete preoperative thyroid ultrasound images with surgical pathology examination | Failure to meet American Thyroid Association criteria for lesions or nodules        | Training dataset: NR (NR; NR) Test dataset: 57 (16; NR) | Training dataset: NR Test dataset: 90% | NR                                                     |
| Stoffel et al (2018)85        | Breast cancer                                                                 | Ultrasound scan and histologically confirmed phyllodes tumour and fibroadenoma    | NR                                                                                 | 34 (NR; NR)                 | NR                               | NR                                                     |
| Streba et al (2012)86         | Hepatological cancer                                                          | Patients with suspected liver masses (with hepatocellular carcinoma, hypervascular and hypovascular liver metastases, hepatic haemangiomas, or focal fatty changes) who underwent contrast-enhanced ultrasound | NR                                                                                 | 58 (NR; 29–89)              | 43%                              | NR                                                     |
| Sun et al (2014)87            | Cardiology                                                                    | Patients with paroxysmal atrial fibrillation or persistent atrial fibrillation     | NR                                                                                 | 60 (11; 29–81)              | 45%                              | NR                                                     |
| Tschandl et al (2019)88       | Dermatological cancer                                                         | Lesions that had lack of pigment, availability of at least one clinical close-up image or one dermatoscopic image, and availability of an unequivocal histopathologic report | Mucosal or missing or poor image cases; equivocal histopathologic reports, cases with <10 examples in the training set category | NR                          | NR                               | NR                                                     |
| Urakawa et al (2019)89        | Trauma and orthopaedics                                                       | All consecutive patients with intertrochanteric hip fractures, and anterior x-ray with compression hip screws | Pseudarthrosis after femoral neck fracture or x-rays showing artificial objects in situ | 85 (NR; 29–104)             | 84%                              | NR                                                     |
| van Grinsven et al (2016)80   | Ophthalmology                                                                 | NR                                                                                 | NR                                                                                 | NR                          | NR                               | NR                                                     |
| Walsh et al (2018)81          | Respiratory disease                                                           | High-resolution CT showing diffuse fibrotic lung disease confirmed by at least two thoracic radiologists | Contrast-enhanced CT                                                               | NR                          | NR                               | NR                                                     |
| Wang et al (2017)82           | Lung cancer                                                                   | PET/CT scan in lobectomy patients with systematic hilar and mediastinal lymph node dissection | NR                                                                                 | 61 (NR; 38–81)              | 46%                              | NR                                                     |

(Table 1 continues on next page)
To estimate the accuracy of deep learning algorithms compared with health-care professionals, we did a subanalysis for studies providing contingency tables for both health-care professional and deep learning algorithm performance tested using the same out-of-sample external validation datasets. Additionally, to address the possibility of dependency between different classification tasks done by the same deep learning algorithm or health-care professional within a study, we did a further analysis on the same studies selecting the single contingency table reporting the highest accuracy for each (calculated as proportion of correct classifications).

As an exploratory analysis, we also pooled performances of health-care professionals and deep learning algorithms derived from internally validated test samples. As with the externally validated results, we selected a single contingency table for each study reporting the highest accuracy for health-care professionals and deep learning algorithms. The purpose of this analysis was to explore whether diagnostic accuracy is overestimated in internal validation alone.

Analysis was done using the Stata 14.2 statistics software package. This study is registered with PROSPERO, CRD42018091176.

Role of the funding source
There was no funding source for this study. The lead authors (XL, LF) and senior author (AKD) had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results
Our search identified 31587 records, of which 20530 were screened (figure 1). 122 full-text articles were assessed for eligibility and 82 studies were included in the systematic review.16–22,24–101 These studies described 147 patient cohorts and considered ophthalmic disease (18 studies), breast cancer (ten studies), trauma and orthopaedics (ten studies), dermatological cancer (nine studies), lung cancer (seven studies), respiratory disease (eight studies), gastroenterological or hepatological cancers (five studies), thyroid cancer (four studies), gastroenterology and hepatology (two studies), cardiology (two studies),...
| Target condition       | Reference standard                                      | Same method for assessing reference standard across samples | Type of internal validation              | External validation |
|------------------------|---------------------------------------------------------|-------------------------------------------------------------|------------------------------------------|---------------------|
| Abbasi-Sureshjani et al (2018) | Diabetes                                                 | Laboratory testing                                         | Yes                                       | Random split sample validation  | No |
| Adams et al (2019)     | Hip fracture                                             | Surgical confirmation                                      | Yes                                       | Random split sample validation  | No |
| Ardila et al (2019)    | Lung cancer                                              | Histology; follow-up                                       | No                                        | NR                   | Yes |
| Ariji et al (2019)     | Lymph node metastasis                                    | Histology                                                  | Yes                                       | Resampling method       | No |
| Ayed et al (2015)      | Breast tumour                                            | Histology                                                  | Yes                                       | Random split sample validation  | No |
| Becker et al (2017)    | Breast tumour                                            | Histology; follow-up                                       | No                                        | Study 1: NA            | Study 2: temporal split-sample validation  | Yes |
| Becker et al (2018)    | Breast tumour                                            | Histology; follow-up                                       | No                                        | Random split sample validation  | No |
| Bien et al (2018)      | Knee injuries                                            | Expert consensus                                           | Internal validation dataset: yes          | External validation dataset: NR | No |
| Brinker et al (2019)   | Melanoma                                                 | Histology                                                  | Yes                                       | Random split sample validation  | No |
| Brown et al (2018)     | Retinopathy                                              | Expert consensus                                           | Yes                                       | Resampling method       | No |
| Burlina et al (2017)   | Age-related macular degeneration                         | Expert consensus                                           | Yes                                       | Resampling method       | No |
| Burlina et al (2018)   | Age-related macular degeneration                         | Reading centre grader                                      | Yes                                       | NR                   | No |
| Burlina et al (2018)   | Age-related macular degeneration                         | Reading centre grader                                      | Yes                                       | NR                   | No |
| Bya et al (2019)       | Breast tumour                                            | Histology; follow-up                                       | No                                        | Resampling method       | Yes |
| Cao et al (2019)       | Prostate cancer                                          | Histology; clinical care notes or imaging reports          | Yes                                       | Resampling method       | No |
| Chee et al (2019)      | Femoral head osteonecrosis                               | Clinical care notes or imaging reports                     | Yes                                       | NR                   | Yes |
| Choi et al (2019)      | Breast tumour                                            | Histology; follow-up                                       | No                                        | NA                   | Yes |
| Choi et al (2018)      | Liver fibrosis                                           | Histology                                                  | Yes                                       | Resampling method       | Yes |
| Ciompi et al (2017)    | Lung cancer                                              | Expert consensus                                           | Yes                                       | Random split sample validation  | Yes |
| Codella et al (2017)   | Melanoma                                                 | Histology                                                  | No                                        | Random split sample validation  | No |
| Coudray et al (2018)   | Lung cancer                                              | Histology                                                  | Yes                                       | NR                   | Yes |
| De Fauw et al (2018)   | Retinal disease                                          | Follow-up                                                  | Yes                                       | Random split sample validation  | No |
| Ding et al (2019)      | Alzheimer's disease                                      | Follow-up                                                  | No                                        | NR                   | Yes |
| Dunnenmon et al (2019) | Lung conditions                                          | Expert consensus                                           | Yes                                       | Resampling method       | No |
| Ehteshami Bejnordi et al (2017) | Lymph node metastases                                  | Histology                                                  | No                                        | Random split sample validation  | Yes |
| Esteva et al (2017)    | Dermatological cancer                                    | Histology                                                  | No                                        | Resampling method       | No |
| Fujioka et al (2019)   | Breast tumour                                            | Histology; follow-up                                       | No                                        | NR                   | No |
| Fujisawa et al (2019)  | Dermatological cancer                                    | Histology                                                  | No                                        | Resampling method       | No |
| Gómez-Valverde et al (2019) | Glaucoma                                             | Expert consensus                                           | Yes                                       | Resampling method       | No |
| Grewal et al (2018)    | Brain haemorrhage                                        | Expert consensus                                           | Yes                                       | NR                   | No |
| Haenssle et al (2018)  | Melanoma                                                 | Histology; follow-up                                       | No                                        | NR                   | No |
| Hamm et al (2019)      | Liver tumour                                             | Clinical care notes or imaging reports                     | Yes                                       | Resampling method       | No |
| Han et al (2018)       | Onychomycosis                                            | Histology; expert opinion on photography                   | No                                        | Random split sample validation  | Yes |
| Han et al (2018)       | Skin disease                                             | Histology; follow-up                                       | No                                        | Random split sample validation  | Yes |
| Hwang et al (2018)     | Pulmonary tuberculosis                                   | Laboratory testing; expert opinion                         | Yes                                       | NR                   | Yes |

(Table 2 continues on next page)
| Target condition | Reference standard | Same method for assessing reference standard across samples | Type of internal validation | External validation |
|------------------|-------------------|----------------------------------------------------------|-----------------------------|---------------------|
| Articles | | | | |

(Continued from previous page)

Hwang et al (2019) 76  Age-related macular degeneration  Expert consensus  Yes  Random split sample validation  Yes

Hwang et al (2019) 77  Lung conditions  Expert consensus  Yes  Random split sample validation  No

Kermany et al (2018) 78  Retinal diseases  OCT: consensus involving experts and non-experts; X-ray: expert consensus  No  Random split sample validation  No

Kim et al (2012) 79  Breast cancer  Histology  Yes  Random split sample validation  No

Kim et al (2018) 80  Maxillary sinusitis  Histology; laboratory testing  Yes  Resampling method  No

Kim et al (2019) 81  Spondylitis  Expert consensus; another imaging modality  Yes  NR  Yes

Kise et al (2019) 82  Sjogren’s syndrome  Expert consensus  Yes  NR  No

Ko et al (2019) 83  Thyroid cancer  Histology  Yes  Resampling method  No

Kumagai et al (2019) 84  Oesophageal cancer  Histology  Yes  NR  No

Lee et al (2019) 85  Intracranial haemorrhage  Expert consensus  Yes  Random split sample validation  Yes

Li C et al (2018) 86  Nasopharyngeal malignancy  Histology  Yes  Random split sample validation  Yes

Li X et al (2019) 87  Thyroid cancer  Histology  Yes  NR  Yes

Lin et al (2014) 88  Breast tumour  Histology  Yes  NR  No

Lindsey et al (2018) 89  Trauma and orthopaedics  Expert consensus  Yes  NR  Yes

Long et al (2017) 90  Ophthalmology  Expert consensus  Yes  Resampling method  Yes

Lu W et al (2018) 91  Macular pathology  Expert consensus  Yes  Resampling method  No

Matsuba et al (2019) 92  Age-related macular degeneration  Expert consensus  Yes  NR  No

Nakagawa et al (2019) 93  Oesophageal cancer  Histology  Yes  NR  Yes

Nam et al (2019) 94  Lung cancer  Expert consensus; another imaging modality, clinical notes  No  Random split sample validation  Yes

Olczak et al (2017) 95  Fractures  Clinical care notes or imaging reports  Yes  Random split sample validation  No

Peng et al (2019) 96  Age-related macular degeneration  Reading centre grader  Yes  NR  No

PoodjASTOt et al (2018) 97  Odontogenic tumours of the jaw  Histology  Yes  NR  No

Rapurkar et al (2018) 98  Lung conditions  Expert consensus  Yes  NR  No

Raumviboonsuk et al (2019) 99  Diabetic retinopathy  Expert consensus  Yes  NR  Yes

Sayes et al (2019) 100  Diabetic retinopathy  Expert consensus  Yes  NR  No

Schlegl et al (2018) 101  Macular diseases  Expert consensus  Yes  Resampling method  No

Shibutani et al (2019) 102  Myocardial stress defect  Expert consensus  Yes  NR  Yes

Shichijo et al (2017) 103  Helicobacter pylori gastritis  Standard-of-care diagnosis based on laboratory testing  No  Random split sample validation  No

Singh et al (2018) 104  Lung conditions  Clinical care notes or imaging reports; existing labels in open-access data library  No  NR  No

Song et al (2019) 105  Thyroid cancer  Histology  Yes  Resampling method  No

Stoffel et al (2018) 106  Breast tumours  Histology  Yes  Random split sample validation  No

Streba et al (2012) 107  Liver tumours  Another imaging modality; histology; follow-up  No  Resampling method  No

(Continued on next page)
oral cancer (two studies), nephrology (one study), neurology (one study), maxillofacial surgery (one study), rheumatology (one study), nasopharyngeal cancer (one study), and urological disease (one study; table 1). One study included two different target conditions. Study characteristics are summarised in the tables (tables 1, 2, 3).

72 studies used retrospectively collected data and ten used prospectively collected data (table 3). 25 studies used data from open-access repositories. No studies reported a prespecified sample size calculation. 26 studies reported that low-quality images were excluded, 18 did not exclude low-quality images, and 38 did not report this. Four studies also tested the scenario where health-care professionals are given additional clinical information alongside the image, and one study tested single image versus the addition of historical images for both health-care professionals and the deep learning algorithm. Four studies also considered diagnostic performance in an algorithm-plus-clinician scenario. Reference standards were wide ranging in line with variation of the target condition and the modality of imaging being used, with some studies adopting multiple methods (table 2). 37 studies used histopathology; 28 studies used varying models of expert consensus; one study relied on single expert consensus; nine studies used clinical follow-up; two studies used surgical confirmation; three studies used reading centre labels (such as when clinical trial data were used); eight studies used existing clinical care notes or imaging reports or existing labels associated with open data sources. Four studies used another imaging modality to confirm the diagnosis and four studies used laboratory testing.

69 studies provided sufficient information to enable calculation of contingency tables and calculation of test performance parameters, with a total of 595 tables across these studies. Within this group, sensitivity for deep learning models ranged from 9·7% to 100·0% (mean 79·1%, SD 0·2) and specificity ranged from 38·9% to 100·0% (mean 88·3%, SD 0·1).

Of the 69 studies, 25 studies did an out-of-sample external validation and were therefore included in a meta-analysis. In line with the aims of this review, all eligible studies were included regardless of the target condition. The meta-analysis therefore included diagnostic classifications in multiple specialty areas, including ophthalmology (six studies), breast cancer (three studies), lung cancer (two studies), dermatological cancer (three studies), trauma and orthopaedics (two studies), respiratory disease (two studies),

| Target condition         | Reference standard | Same method for assessing reference standard across samples | Type of internal validation | External validation |
|--------------------------|--------------------|-----------------------------------------------------------|----------------------------|---------------------|
| Sun et al (2014)          | Atrial thrombi     | Surgical confirmation, another imaging modality, clinical care notes or imaging reports | No                         | Random split sample validation | No                 |
| Tschandl et al (2019)     | Dermatological cancer | Histology                                              | Yes                        | NR                  | Yes                |
| Uraikawa et al (2019)     | Hip fractures      | Clinical care notes or imaging reports                  | Yes                        | Random split sample validation | No                 |
| van Grinsven et al (2016) | Retinal haemorrhage| Single expert                                            | Yes                        | Random split sample validation | Yes                |
| Walsh et al (2018)        | Lung fibrosis      | Expert consensus                                         | Yes                        | NR                  | Yes                |
| Wang et al (2017)         | Lymph node metastasis | Expert consensus                                    | Yes                        | Resampling method    | No                 |
| Wang et al (2018)         | Lung cancer        | Histology                                                | Yes                        | Random split sample validation | No                 |
| Wang et al (2019)         | Malignant thyroid nodule | Histology                                                | Yes                        | NR                  | No                 |
| Wright et al (2014)       | Renal tissue function | Clinical care notes or imaging reports                   | Yes                        | Random split sample validation | No                 |
| Wu et al (2019)           | Gastric function   | Histology                                                | Yes                        | Resampling method    | No                 |
| Ye et al (2019)           | Intracranial haemorrhage | Expert consensus                                    | Yes                        | Random split sample validation | No                 |
| Yu et al (2018)           | Melanoma           | Histology                                                | Yes                        | Resampling method    | No                 |
| Zhang C et al (2019)      | Lung cancer        | Expert consensus                                         | Yes                        | Resampling method    | Yes                |
| Zhang Y et al (2019)      | Retinopathy        | Expert consensus                                         | Yes                        | Random split sample validation | No                 |
| Zhao et al (2018)         | Lung cancer        | Histology                                                | Yes                        | NR                  | No                 |

Blinded assessment of reference standard was not reported in any of the studies. NR=not reported. OCT=optical coherence tomography. DMSA=2,3-dimercapto-succinic acid.

Table 2: Model training and validation for the 82 included studies
| Indicator definition | Algorithm | Data source |
|----------------------|-----------|-------------|
| Fundus image         | ResNet    | Retrospective cohort; secondary analysis of a subset of the Maastricht’s study—a population-based cohort (collected in the southern part of the Netherlands), enriched with patients with diabetes |
| X-ray                | AlexNet   | Retrospective cohort; data from the Royal Melbourne Hospital (Melbourne, VIC, Australia) radiographic archive |
| CT                   | Mask RNN, RetinaNet, Inception V1 | Retrospective clinical trial data from the National Lung Cancer Screening Trial |
| Contrast-enhanced CT | CNN       | Retrospective cohort; data from the Aichi-Gakuin University School of Dentistry (Nagoya, Japan) |
| Mammograms           | ANN       | Retrospective cohort; secondary analysis of a subset of the Farabi Digital Database for Screening Mammography collected at the radiology centre ElFarabi (Tunisia) |
| Breast ultrasound    | ViDi Suite Version 20 | Retrospective cohort; data collected at the University Hospital Zurich (Zurich, Switzerland) |
| MRI                  | MRNet     | Retrospective cohort; data from the Stanford University Medical Center (CA, USA) |
| Dermoscopy           | ResNet-50 | Retrospective cohort; data collected at multiple institutions for a research challenge (International Skin Image Collaboration) |
| Fundus image         | CNN       | Retrospective cohort; data collected at multiple hospitals across North America |
| Fundus image         | AlexNet, OverFeat | Retrospective cohort; secondary analysis of a subset from the Age-related Eye Disease Study trial |
| Fundus image         | ResNet    | Retrospective cohort; secondary analysis of a subset from the Age-related Eye Disease Study trial |

(Table 3 continued on next page)
| Indicator definition | Algorithm architecture name | Algorithm architecture | Data source | Number of images for training/tuning | Source of data | Data range | Open-access data |
|----------------------|-----------------------------|------------------------|-------------|-------------------------------------|----------------|------------|----------------|----------------|
| Byra et al (2019)²⁸ | Ultrasound                   | No                     | No          | VGG-19                              | CNN; VGG      | 512/150    | Three data cohorts: Cohort 1: retrospective cohort; secondary analysis from the Moores Cancer Center, University of California (San Diego, CA, USA) Cohort 2: retrospective cohort; secondary analysis of the UOCAT Diagnostic Centre of the Parc Tauli Corporation (Sabadell, Spain) Cohort 3: Retrospective cohort; data collected from the Institute of Oncology (Warsaw, Poland) |              |
| Cao et al (2019)³⁵  | MRI                          | NR                     | No          | FocalNet                            | CNN           | NR/NR      | Retrospective cohort | NR            | No             |
| Choi et al (2019)³⁷ | Ultrasound                   | Yes                    | No          | GoogleNet                           | CNN; Inception| 790/NR     | Retrospective cohort; secondary analysis of data from Samsung Medical Center (Seoul, Korea) | 2015          | No             |
| Choi et al (2018)³⁰ | Contrast-enhanced CT         | NR                     | No          | CNN                                 | CNN           | 7461/NR    | Three data cohorts: Cohort 1: retrospective cohort; secondary analysis of data from Asan Medical Center (Seoul, South Korea) used for development dataset Cohort 2: retrospective cohort; data collated from LG University Paik Hospital and Hanyang University Hospital (both Seoul, South Korea) for external validation Cohort 3: retrospective cohort; data from Yonsei University Severance Hospital (Seoul, South Korea) |              |
| Ciompi et al (2017)³⁹ | CT                           | No                     | No          | ConvNets                            | CNN           | 490 320/453| Retrospective cohort; secondary analysis of a subset of the multicentre Italian Lung Detection trial and the Danish Lung Cancer Screening Trial | NR            | Yes            |
| Codella et al (2017)⁴⁰ | Histology                   | NR                     | No          | MA (ensembles)                      | NR            | NR/NR      | Retrospective cohort; data collected at multiple institutions for a research challenge (International Skin Imaging Collaboration) | NR            | Yes            |
| Coudray et al (2018)⁴¹ | Histology                   | No                     | No          | Inception V3                        | CNN; Inception| 825/148    | Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of a subset from the Cancer Genome Atlas Program, National Cancer Institute Cohort 2: retrospective cohort; secondary analysis of data from the NewYork University Langone Medical Center (New York, NY, USA) for external validation | Cohort 1: NR | Cohort 2: NR   | Yes            |
| De Fauw et al (2018)⁴² | OCT                          | Yes                    | No          | 3D U-Net                            | CNN; U-Net    | 14884/933  | Retrospective cohort; secondary analysis of a subset of data collected at Moorfields Eye Hospital (London, UK) | 2012–17       | No             |

(Continued from previous page)
| Indicator definition | Algorithm | Data source |
|----------------------|-----------|------------|
| Method for predictor measurement | Exclusion of poor-quality imaging | Heatmap provided |
| Algorithm architecture name | Algorithm architecture | Transfer learning applied | Number of images for training/tuning | Source of data | Data range | Open-access data |
| Exclusion of poor-quality imaging | Heatmap provided | Algorithm architecture | Transfer learning applied | Number of images for training/tuning | Source of data | Data range | Open-access data |

(Continued from previous page)

Ding et al (2019) NR PET Yes Inception V3 CNN; Inception Yes 1921/NR Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of a prospectively collected dataset accessible to the Alzheimer’s Disease Neuroimaging Initiative Cohort 2: retrospective cohort; secondary analysis of data from the Department of Radiology, University of California (CA, USA) for external validation

Dunnimon et al (2019) NR Chest x-ray Yes ResNet-18, Dense-Net 121, K SVM + BOW; AlexNet CNN; Residual Network; DenseNet Support Vector Machine Yes 180 000/20 000 Retrospective cohort; secondary analysis of data from the Department of Radiology, Stanford University (CA, USA)

Ehteshami Bejnordi et al (2017) NR Histology No Yes GoogleLeNet; ResNet; VGG-16, VGG-Net; SegNet; U-Net; CRResNet; 7-layer CNN, 3-layer VGG CNN; Inception; VGG; Residual Network; AlexNet; U-Net Yes 270/NR Retrospective cohort; data collected at the Radboud University Medical Center (Nijmegen, Netherlands) and University Medical Center (Utrecht, Netherlands) for a research challenge (“CAMELYON16”) 2015 Yes

Esteva et al (2017) NR Photographs Yes Yes Inception V3 CNN; Inception Yes 129 450/NR Retrospective cohort; includes data from online open-access repositories (ie, the International Skin Imaging Collaboration Dermoscopic Archive, the Edinburgh Dermofit Library) and data from Stanford Hospital (CA, USA) NR Yes

Fujio et al (2019) NR Breast ultrasound NR No Inception V2 CNN; Inception Yes 947/NR Retrospective cohort; secondary analysis of data from Tokyo Medical and Dental University (Tokyo, Japan) 2010–17 No

Fujisawa et al (2019) NR Photographs NR No GoogleNet CNN; Inception Yes 4867/NR Retrospective cohort; secondary analysis of data from University of Tsukuba Hospital (Tsukuba, Japan) 2003–16 No

Gómez-Vaquer et al (2019) NR Fundus images Yes No VGG-19 CNN; VGG Yes 1560/NR Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of data collated into the RIM-ONE and DRISHTI-GS datasets Cohort 2: retrospective cohort; secondary analysis of data from Hospital de la Esperanza (Parc de Salut Mar, Barcelona, Spain) NR Yes

Grewal et al (2018) NR CT NR No DenseNet CNN; DenseNet Yes 185 67 Retrospective dataset from two local hospitals NR No

(Table 3 continued on next page)
| Indicator definition | Algorithm architecture | Data source |
|----------------------|------------------------|-------------|
| **Method for predictor measurement** | **Algorithm architecture** | **Source of data** |
| Exclusion of poor-quality imaging | Heatmap provided | Number of images for training/tuning | Data range | Open-access data |
| **Haenssle et al (2018)** | Dermoscopy | NR | No | Inception V4 | CNN | Yes | NR/NR | Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of data from the validated image library at the Department of Dermatology, University of Heidelberg, Germany Cohort 2: retrospective cohort; secondary analysis of a subset of data from the International Skin Imaging Collaboration melanoma project | Cohort 1: NR | Cohort 2: NR | Cohort 1: no | Cohort 2: yes |
| **Hammet et al (2019)** | Contrast-enhanced MRI | Yes | No | CNN | CNN | No | 434/NR | Retrospective cohort; secondary analysis of data collected at the Department of Radiology and Biomedical Imaging, Yale School of Medicine (New Haven, CT, USA) | 2010–17 | No |
| **Han et al (2018)** | Photographs | No | Yes | Ensemble: ResNet-152 + VGG-19 (arithmetic mean of both outputs) | Ensemble, CNN, Residual Network | Yes | 19398/NR | Retrospective cohort; data collected at the Asan Medical Center, Inje, and Seoul University (Seoul, South Korea) and Hallym (Dongtan, South Korea) | 2003–16 | Yes |
| **Han et al (2018)** | Photographs | Yes | No | ResNet-152 | CNN, Residual Network | Yes | 49 567/341 | Retrospective cohort; data collected at the Asan Medical Center (Seoul, South Korea), University Medical Center Groningen (Groningen, Netherlands), Dongtan Sacred Heart Hospital (Gyeonggi, South Korea), Hallym University (Dongtan, South Korea), Sanggye Pahk Hospital (Seoul, South Korea), Inje University (Seoul, South Korea), with additional data collected from open-access repositories and websites | NR | Yes |
| **Hwang et al (2018)** | Chest x-ray | NR | Yes | CNN | CNN | No | 60 689/450 | Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of data collected from the imaging database of Seoul National University Hospital (Seoul, South Korea) Cohort 2: retrospective cohort; secondary analysis of data collected at Seoul National University Hospital (Seoul, South Korea), Boramae Medical Center (Seoul, South Korea), Kyunghee University Hospital at Gangdong (Seoul, South Korea), and Daejeon Eulji Medical Center (Daejeon, South Korea) Cohort 3: retrospective cohort; secondary analysis of the tuberculosis screening programme of Montgomery County (MD, USA) Cohort 4: retrospective cohort; secondary analysis of data from the tuberculosis screening programme of Shenzhen, China | Cohort 1: 2010–16 | Cohort 2: 2016–17 | Cohort 3: NR | Cohort 4: NR | Cohort 1: no | Cohort 2: no | Cohort 3: yes | Cohort 4: yes |

(Table 3 continued on next page)
| Indicator definition | Algorithm | Data source | Number of images for training/tuning | Number of images for testing/validation | Number of images for training/tuning | Number of images for testing/validation | Data range | Open-access data | Method for predictor measurement | Exclusion of poor-quality imaging | Heatmap provided | Algorithm architecture name | Algorithm architecture | Transfer learning applied | Source of data | Data source | Data range | Open-access data |
|----------------------|-----------|-------------|-------------------------------------|----------------------------------------|---------------------------------------|---------------------------------------|------------|----------------|------------------|-----------------------------|------------------------|------------------|-----------------------------|-----------------------------|---------------------------|------------------------|------------------|------------------------|------------------------|
| Hwang et al (2019)   | OCT       | Yes         | VGG-16; Inception V3; ResNet-50     | Yes                                    | Yes                                   | 287207/810                           | Retrospective cohort; secondary analysis of data collected at the Department of Ophthalmology of Taipei Veterans General Hospital (Taipei, Taiwan) | 2017       | No                    |                         |                         |                        | CNN, VGG; Inception, Residual Network | CNN, VGG, Inception, Residual Network | Yes                      |                         | 2017       | No                    |
| Hwang et al (2019)   | Chest x-ray | No          | Yes                                 | CNN                                    | No                                    | 876951050                             | Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of data from a single institution (no further details available) Cohort 2: retrospective cohort; data collated from four hospitals in Korea and four hospitals in France for external validation | 2010–17    | No                    |                         |                         |                        | CNN                                    | CNN                                    | No                      |                         | 2010–17    | No                    |
| Kermany et al (2018) | OCT       | Yes         | Inception V3                        | Yes                                    | Yes                                   | OCT 3121000                           | OCT: retrospective cohort of adult patients from the Shirley Eye Institute of the University of California San Diego (CA, USA), the California Retinal Research Foundation (CA, USA), Medical Center Ophthalmology Associates, Shanghai First People’s Hospital (Shanghai, China), and Beijing Tongren Eye Center (Beijing, China) Chest x-ray: retrospective cohort of adolescent patients aged 1-5 years from Guangzhou Women and Children’s Medical Center (Guangzhou, China) | 2013, 2017 | Yes                    |                         |                         |                        | OCT, Inception V3                                    | CNN                                    | Yes                      |                         | 2013, 2017 | Yes                    |
| Kim et al (2012)     | Ultrasound | No          | No                                  | 3-layer back propagation ANN (multilayered perceptron) | No                                    | 7070                                  | Retrospective cohort; data collected at the Kangwon National University College of Medicine (Gangwon-do, South Korea) | 2001–03    | No                    |                         |                         |                        | ANN                                    | ANN                                    | No                      |                         | 2001–03    | No                    |
| Kim et al (2018)     | Waters’ view plain film radiography | No          | No                                  | CNN                                    | No                                    | NR/NR                                 | Retrospective cohort; secondary analysis of data collected from Gangnam Severance Hospital (Seoul, South Korea) | 2007–16    | No                    |                         |                         |                        | CNN                                    | CNN                                    | No                      |                         | 2007–16    | No                    |
| Kim et al (2019)     | MRI       | No          | Yes                                 | ResidualNet                            | Yes                                   | 80001000                              | Retrospective cohort; secondary analysis of data collected from Seoul National University Hospital and Seoul National University Bundang Hospital (Seoul, South Korea) | 2003–17    | No                    |                         |                         |                        | CNN, Residual Network                                    | CNN, Inception Residual Network | Yes                      |                         | 2003–17    | No                    |
| Kise et al (2019)    | CT        | NR          | No                                  | AlexNet                                | Yes                                   | 400/NR                                | Retrospective cohort                    | NR          | No                    |                         |                         |                        | CNN, AlexNet                                    | CNN, AlexNet | Yes                      |                         | 2012–15    | No                    |
| Ko et al (2019)      | Ultrasound | NR          | No                                  | CNN (imagenet-vgg-verydeep-16 and Imagenet-VGG-F) | Yes                                   | 594/NR                                | Retrospective cohort; secondary analysis of data collected at the Department of Radiology, Jeju National University Hospital, Jeju National School of Medicine (Jeju, South Korea) | 2012–15    | No                    |                         |                         |                        | CNN, VGG                                    | CNN, VGG, Inception Residual Network | Yes                      |                         | 2012–15    | No                    |
| Kumagai et al (2019) | Endoscopy images | Yes          | No                                  | GoogleNet                              | Yes                                   | 240/NR                                | Retrospective cohort; secondary analysis of data collected from Saitama Medical Center (Saitama, Japan) | 2011–18    | No                    |                         |                         |                        | CNN, Inception                                    | CNN, Inception Residual Network, Inception | Yes                      |                         | 2011–18    | No                    |
| Lee et al (2019)     | CT        | No          | Yes                                 | VGG-16, ResNet-50; Inception V3; Inception-ResNet-V2 ensemble | Yes                                   | 704200                                | Retrospective cohort for development; both retrospective and prospective cohort for validation; data collected from Department of Radiology, Massachusetts General Hospital (Boston, MA, USA) | 2003–17    | No                    |                         |                         |                        | CNN, VGG, Inception Residual Network                                    | CNN, VGG, Inception Residual Network | Yes                      |                         | 2003–17    | No                    |

(Continued from previous page)
| Indicator definition | Algorithm architecture name | Algorithm architecture | Transfer learning applied | Number of images for training/tuning | Source of data | Data range | Open-access data |
|----------------------|-----------------------------|------------------------|--------------------------|-------------------------------------|----------------|------------|----------------|
| (Continued from previous page) |
| Li et al (2018) | Endoscopic photographs | Yes | Yes | CNN | CNN | Yes | 5557/807 | Combined retrospective and prospective cohort; data collected from the Sun Yat-sen University Cancer Center (Guangzhou, China) | 2008–17 | No |
| Li et al (2019) | Ultrasound | Yes | Yes | ResNet-50 and Darnet-19, Ensemble DCNN | CNN; Residual Network | Yes | 42952/NR | Retrospective cohort; secondary analysis of data collected from Tianjin Cancer Hospital (Tianjin, China), Integrated Traditional Chinese and Western Medicine Hospital (Jilin, China) and Weihai Municipal Hospital (Shandong, China) | 2012–18 | No |
| Lin et al (2014) | Ultrasound | NR | No | Fuzzy Cerebellar Neural Network | Fuzzy Neural Network | No | NR/NR | Retrospective cohort; data collected at the Far Eastern Memorial Hospital, Taiwan | 2006-07 | No |
| Lindsey et al (2018) | X-ray | NR | Yes | DCNN | U-Net, DCNN | No | 100 535/28 341 | Retrospective cohort; Department of Orthopaedic Surgery, Hospital for Special Surgery (New York, NY, USA) | 2016 | No |
| Long et al (2017) | Ocular photographs | No | No | CC-Cruiser, DONN | DCNN | No | 886/NR | Three data cohorts: Cohort 1: retrospective cohort; Childhood Cataract Program of the Chinese Ministry of Health Cohort 2: prospective cohort; three non-specialised collaborating hospitals in China (two sites in Guangzhou City and one in Qingyuan City) Cohort 3: search engine cohort; image searches of the search engines Google, Baidu, and Bing | Cohort 1: NR | No |
| Lu et al (2018) | OCT | Yes | Yes | ResNet | CNN, Residual Network | Yes | 19 815/2 202 | Retrospective cohort; data collected from Wuhan University Eye Center (Wuhan, China) | 2012–14 | No |
| Matsuba et al (2019) | Fundus images | NR | Yes | CNN | CNN | No | 253/NR | Retrospective cohort; secondary analysis of data collected from Tsukazaki Hospital (Himeji, Japan) | NR | No |
| Nakagawa et al (2019) | Endoscopic images | Yes | No | VGG | VGG | No | 804/NR | Retrospective cohort; secondary analysis of data from Osaka International Cancer Institute (Osaka, Japan) | 2005-18 | No |
| Nam et al (2019) | Chest x-ray | Yes | Yes | DCNN | CNN | No | 43292,600 | Retrospective cohort; secondary analysis of data collected from Seoul National University Hospital (Seoul, South Korea), Boramae Hospital (Seoul, South Korea), and National Cancer Center, University of California San Francisco Medical Center (San Francisco, CA, USA) | 2010–17 | No |
| Olczak et al (2017) | Wrist, hand, and ankle x-rays | NR | No | VGG-16 | CNN, VGG | Yes | 179 527/1NR | Retrospective cohort; data collected at the Danderyd Hospital (Danderyd, Sweden) | 2002–15 | No |
| Peng et al (2019) | Fundus images | No | Yes | Pnet, LA-net, DeepSet Net, Inception V3, CNN, Dnet | CNN, Inception | Yes | 58 402/NR | Retrospective cohort; data collected as part of the AREDS study dataset | 1992–2005 | Yes |
| Poesada-Teoti et al (2018) | X-ray | NR | Yes | VGG-16 | CNN, VGG | Yes | 400/NR | Retrospective cohort; secondary analysis of data collected from Thammasat University (Pathumthani, Thailand) | NR | No |

(Table 3 continued on next page)
| Indicator definition | Algorithm | Data source | Data range | Open-access data |
|----------------------|-----------|-------------|------------|-----------------|
| Method for predictor measurement | Exclusion of poor-quality imaging | Heatmap provided | Algorithm architecture name | Algorithm architecture | Transfer learning applied | Number of images for training/tuning | Source of data | |
| Rajpurkar et al (2018) | Chest x-ray | NR | Yes | DenseNet | CNN; DenseNet | NR | 9863/8351 | Retrospective cohort; secondary analysis of data within the ChestX-ray14 depository of the National Institutes of Health | NR | Yes |
| Raumbooisuk et al (2019) | Fundus images | Yes | No | CNN Inception-V4 | CNN; Inception | NR | Not applicable/not applicable | Retrospective cohort; secondary analysis of data collected from the National Thai Registry of Diabetic Retinopathy | 2015-2017 | No |
| Sayres et al (2019) | Fundus images | Yes | Yes | Inception V4 | CNN; Inception | Yes | 140000/2000 | Retrospective cohort; secondary analysis of data collected from Aravind Eye Hospital (India), Sankara Nethralaya Hospital (Chennai, India), and Narayana Nethralaya Hospital (India) | 2015 | No |
| Schlegl et al (2018) | OCT | Yes | Yes | Deep learning model | CNN | No | 840/NR | Retrospective cohort; secondary analysis of data collected from the Vienna Reading Center, Vienna, Austria | NR | No |
| Shibutani et al (2019) | SPECT | NR | Yes | ANN | ANN | No | Not applicable/not applicable | Prospective secondary analysis of data collected from Kanazawa University Hospital, Japan | NR | No |
| Shichijo et al (2017) | Endoscopic images | Yes | No | GoogleNet and Caffe DL framework | CNN; Inception | Yes | 32203/NR | Retrospective cohort; data collected at Tada Toromihro Institute of Gastroenterology and Pictology, Saitama, Japan | 2014–16 | No |
| Singh et al (2018) | Chest x-ray | No | Yes | “Qure AI” | CNN | No | 115008/93972 | Two data cohorts: Cohort 1: retrospective cohort for training “various hospitals” in India Cohort 2: retrospective cohort for testing: ChestX-ray8 database | Cohort 1: NR | Cohort 2: NR |
| Song et al (2019) | Ultrasound | NR | No | Multitask cascade pyramid CNN | CNN | Yes | 6228/NR | Two data cohorts: Cohort 1: retrospective cohort for training “various hospitals” in India Cohort 2: retrospective cohort from open repository | Cohort 1: NR | Cohort 2: NR |
| Stoffel et al (2018) | Ultrasound | NR | No | Proprietary ViDi Suite deep-learning system | CNN | No | 53/NR | Retrospective cohort; secondary analysis of data from the University Hospital Zurich, Switzerland | 2013-15 | No |
| Streba et al (2012) | Contrast-enhanced ultrasound | NR | No | Multilayer ANN | ANN | No | NR/NR | Prospective cohort of patients that underwent contrast-enhanced ultrasound imaging collected at the Research Center of Gastroenterology and Hepatology, Craiova, Romania | 2008-11 | No |
| Sun et al (2014) | Transoesophageal echocardiogram | NR | No | ANN | ANN | No | NR/NR | Prospective cohort of patients collected at the Hospital of Harbin Medical University (Harbin, China) | 2006-11 | No |
| Tschandl et al (2019) | Dermoscopy and clinical close-up | Yes | No | Inception V3 | CNN; Inception | No | 13724975 | Two data cohorts: Cohort 1: retrospective cohort; secondary analysis of data from two open-source repositories (EDRA and ISIC) Cohort 2: retrospective cohort; secondary analysis of data from a single skin cancer centre | Cohort 1: NR | Cohort 2: 2001-16 | 2001-16 | No | Yes |

(Continued from previous page)
| Indicator definition | Algorithm | Data source |
|----------------------|-----------|-------------|
| Exclusion of poor-quality imaging | Algorithm architecture name | Algorithm architecture |
| Heatmap provided | Transfer learning applied | Number of images for training/tuning |
| Source of data | Data range | Open-access data |

(Continued from previous page)

| Urakawa et al (2019) | X-ray | NR | No | VGG-16 | CNN, VGG | Yes | 2978/334 | Retrospective cohort; secondary analysis of data collected from Department of Orthopedic Surgery, Tsuruoka, Municipal Shonai Hospital (Japan) | 2006–17 | No |
| van Grinsven et al (2016) | Fundus images | NR | Yes | SES CNN 60; NSES CNN 170 | CNN | No | 5287/NR | Retrospective cohort; secondary analysis of two open-access online repositories | NR | Yes |
| Walsh et al (2018) | High-resolution CT | NR | No | Inception-ResNet-V2 | CNN, Inception | Yes | 929/89 | Retrospective cohort; secondary analysis of data from La Fondazione Policlinico Universitario A Gemelli IRCCS (Rome, Italy) and University of Parma (Parma, Italy) | NR | No |
| Wang et al (2017) | PET/CT | NR | No | BP-ANN D13; BP-ANN T82; BP-ANN A95; BP-ANN S6; CNN | ANN | No | NR/NR | Retrospective cohort; data collected at the Affiliated Tumor Hospital of Harbin Medical University (Harbin, China) | 2009–14 | No |
| Wang et al (2018) | High-resolution CT | NR | No | 3DCNN | CNN | No | 1075/270 | Retrospective cohort; secondary analysis of data collected from Fudan University Shanghai Cancer Center (Shanghai, China) | 2010–17 | No |
| Wang et al (2019) | Ultrasound | NR | No | ResNet V2-50; YOLOv2 | CNN, Residual Network | Yes | 5007/NR | Retrospective cohort; secondary analysis of data from Affiliated Hospital of Qingdao University (China) | 2018 | No |
| Wright et al (2014) | Dimercaptosuccinic acid images | Yes | No | CNN | CNN | No | 257/NR | Retrospective cohort; data collected consecutively across two sites | 2009–11 | No |
| Wu et al (2019) | Endoscopy images | Yes | Yes | VGG-16; ResNet-50 | CNN, VGG, Residual Network | Yes | 24549/NR | Retrospective cohort; secondary analysis of data from Remin Hospital of Wuhan University (Wuhan, China) | 2018 | No |
| Ye et al (2019) | CT | Yes | Yes | CNN, RNN | CNN, RNN | No | 2255/282 | Retrospective cohort; secondary analysis of data collected from three participating hospitals | 2013–18 | No |
| Yu et al (2018) | Dermoscopy | NR | No | CNN, VGG-16 | CNN, VGG | Yes | 362/109 | Retrospective cohort; secondary analysis of data from Severance Hospital in the Yonsei University Health System (Seoul, South Korea) and Dongsan Hospital in the Keimyung University Health System (Daegu, South Korea) | 2013–16 | No |
| Zhang C et al (2019) | CT | No | No | CNN | CNN | No | 2285/757 | Retrospective cohort; secondary analysis of data from Guangdong Provincial People’s Hospital, Foshan First People’s Hospital, and Guangdong Lung Cancer Institute (Guangzhou, China) and The Third Affiliated Hospital of Sun Yat-sen University and Guangzhou Chest Hospital (Guangzhou, China) | 2015-17 | Yes |
| Zhang Y et al (2019) | Fundus images | Yes | No | VGG-16 | CNN, VGG | Yes | 17801/NR | Retrospective cohort; secondary analysis of data from Shenzhen Eye Hospital regional screening data (Shenzhen, China) | NR-2018 | No |
| Zhao et al (2018) | CT | NR | No | DenseNet | CNN, DenseNet | No | 523/NR | Retrospective cohort; secondary analysis of data from Huadong Hospital Affiliated to Fudan University (Shanghai, China) | 2011-17 | No |

Number and type of predictors was not reported in any of the studies. NR=not reported. CNN=convolutional neural network. ANN=artificial neural network. RNN=recurrent neural network. DCNN=deep CNN. OCT=optical coherence tomography.

Table 3: Indicator, algorithm, and data source for the 82 included studies
and one study each for cardiology, gastroenterology or hepatology, gastroenterological or hepatological cancer, maxillofacial surgery, thyroid cancer, neurology, and nasopharyngeal cancer. These studies included 141 patient cohorts. Six studies included prospectively collected data, whereas all others used retrospective data. Nine studies used data from open-access repositories. In total, 161 contingency tables were included in the meta-analysis (appendix pp 3–6).

Hierarchical summary ROC curves of these 25 studies (161 contingency tables) are shown in figure 2. When averaging across studies, the pooled sensitivity was 88·6% (95% CI 85·7–90·9) for all deep learning algorithms and 79·4% (74·9–83·2) for all health-care professionals. The pooled specificity was 93·9% (92·2–95·3) for deep learning algorithms and 88·1% (82·8–91·9) for health-care professionals.

Of these 25 studies, only 14 used the same sample for the out-of-sample validation to compare performance between deep learning algorithms and health-care professionals, with 31 contingency tables for deep learning algorithm performance and 54 tables for health-care professionals (figure 3). The pooled sensitivity was 85·7% (95% CI 78·6–90·7) for deep learning algorithms and 79·4% (74·9–83·2) for health-care professionals. The pooled specificity was 93·5% (89·5–96·1) for deep learning algorithms and 87·5% (81·8–91·6) for health-care professionals.

After selecting the contingency table reporting the highest accuracy for each of these 14 studies (ie, 14 tables for deep learning algorithms and 14 tables for health-care professionals), the pooled sensitivity was 87·0% (95% CI 83·0–90·2) for deep learning algorithms and 86·4% (79·9–91·0) for health-care professionals. The pooled specificity was 92·5% (85·1–96·4) for deep learning algorithms and 90·5% (80·6–95·7) for health-care professionals (figure 4).

As an exploratory analysis, we also pooled performances of health-care professional and deep learning algorithms derived from matched internally validated samples (37 studies). Again, we selected a single contingency table for each study reporting the highest accuracy. In this sample, all accuracy metrics were higher, with a pooled sensitivity of 90·1% (95% CI 86·9–92·6) for deep learning algorithms and 90·5% (86·3–93·5) for health-care professionals and a pooled specificity of 93·3% (90·1–95·6) for deep learning algorithms and 91·9% (87·8–94·7) for health-care professionals (figure 4).

Discussion
To our knowledge, this is the first systematic review and meta-analysis on the diagnostic accuracy of health-care professionals versus deep learning algorithms using medical imaging. After careful selection of studies with transparent reporting of diagnostic performance and validation of the algorithm in an out-of-sample population, we found deep learning algorithms to have equivalent sensitivity and specificity to health-care professionals. Although this estimate seems to support the claim that deep learning algorithms can match clinician-level accuracy, several methodological deficiencies that were common across most included studies should be considered.

First, most studies took the approach of assessing deep learning diagnostic accuracy in isolation, in a way that does not reflect clinical practice. Many studies were excluded at screening because they did not provide comparisons with health-care professionals (ie, human vs machine), and very few of the included studies reported comparisons with health-care professionals using the same test dataset. Considering deep learning algorithms in this isolated manner limits our ability to extrapolate the findings to health-care delivery, except perhaps for
mass screening. Only four studies provided health-care professionals with additional clinical information, as they would have in clinical practice; one study also tested the algorithm-plus-clinician scenario in which prior or historical imaging was provided to the algorithm and the health-care professional, and four studies also considered diagnostic performance in an algorithm-plus-clinician scenario. It is worth noting that no studies reported a formal sample size calculation to ensure that the study was sufficiently sized in a head-to-head comparison. Although we acknowledge that sample size calculations can be challenging in this context, a lack of consensus on principled methods to perform them is no justification to ignore them in the design of a study.

Second, there were very few prospective studies done in real clinical environments. Most studies were retrospective, in silico, and based on previously assembled datasets. The ground truth labels were mostly derived from data collected for other purposes, such as in retrospectively collected routine clinical care notes or radiology or histology reports, and the criteria for the presence or absence of disease were often poorly defined. The reporting around handling of missing information in these datasets was also poor across all studies. Most did not report whether any data were missing, what proportion this represented and how missing data were dealt with in the analysis. Such studies should be considered as hypothesis generating, with real accuracy defined in patients, not just datasets.

Third, a wide range of metrics were employed to report diagnostic performance in deep learning studies. If a probability function is not reported, the frequency of true positives, false positives, false negatives, and true negatives at a specified threshold should be the minimum requirement for such comparisons. In our review, only 12 studies reported the threshold at which sensitivity and specificity were reported, without justification of how the threshold was chosen; choice of threshold is often set at the arbitrary value of 0.5, as is convention in machine learning development. Metrics commonly used in the field of computer science, such as accuracy, precision, dice coefficient, and F1 score, are sometimes the only measure for reporting diagnostic performance. Since these tests are usually performed at a prevalence of 50%, these parameters are less comprehensive and useful for clinical practice.

Fourth, there is inconsistency over key terminology used in deep learning studies. Distinct datasets with independent samples should be defined in the development of a deep learning model from the initial training set through to one or more test sets that support validation. We found that the term “validation” is used variably, with some authors using the term appropriately for testing of the final model but others using it for the tuning of a model during development. It is crucial that the validation test set contains data independent to training or tuning data and is used only for assessing the final model. In several studies, we found a lack of transparency as to whether the test set was truly independent due to this inconsistent use of terminology. A standard nomenclature should be adopted. We suggest distinguishing the datasets involved in the development of an algorithm as training set (for training the algorithm), tuning set (for tuning hyperparameters), and validation test set (for estimating the performance of the algorithm). For describing the different types of validation test sets, we suggest adoption of the suggestion by Altman and Royston: internal validation (for in-sample validation), temporal validation (for in-sample validation with a temporal split), and external validation (for out-of-sample validation).

Finally, although most studies did undertake an out-of-sample validation, most did not do this for both health-care professionals and deep learning algorithms. Moreover, only a small number of studies tested the performance of health-care professionals and deep learning algorithms in the same sample. In this review, we accepted both geographically and temporally split
test data, as well as the use of open-access datasets, as external validations. For internal validation, most studies adopted the approach of randomly splitting a single sample into training, tuning, and test sets, instead of preferred approaches such as resampling methods (eg, bootstrapping and cross validation), which have been recommended in clinical prediction model guidelines.9

Our finding when comparing performance on internal versus external validation was that, as expected, internal validation overestimates diagnostic accuracy in both health-care professionals and deep learning algorithms. This finding highlights the need for out-of-sample external validation in all predictive models.

An encouraging finding of this review is the improvement in quality of studies within the last year. 58 (71%) of the 82 studies satisfying the inclusion criteria were newly identified in the updated search, suggesting that the past year has seen a substantial increase in the number of studies comparing algorithm accuracy with health-care professionals. Only five studies additionally did external validation for algorithms and health-care professionals and were eligible for meta-analysis before the updated search, whereas a further 20 studies were suitable for meta-analysis in the review update. A persistent problem is studies not reporting contingency tables (or of sufficient detail for construction of contingency tables), as we were unable to construct contingency tables for two (9%) of 22 studies in the original search and 11 (18%) of 60 studies in the updated search.

Our final comparison estimating the differences in diagnostic accuracy performance between deep learning algorithms and health-care professionals is based on a relatively small number of studies. Less than a third of the included studies were eligible for meta-analysis. This is a direct consequence of poor reporting and lack of external validation in many studies, which has resulted in inadequate data availability and thus exclusion from the meta-analysis. We acknowledge that inadequate reporting does not necessarily mean that the study itself was poorly designed and, equally, that poor study design does not necessarily mean that the deep learning algorithm is of poor quality. Accordingly, there is considerable uncertainty around the estimates of diagnostic performance provided in our exploratory meta-analysis and we must emphasise that reliable estimates of the level of performance can only be achieved through well designed and well executed studies that minimise bias and are thoroughly and transparently reported.

We have not provided a systematic quality assessment for transparency of reporting in this review. This decision was made because existing reporting guidelines for prediction models, such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement, are focused primarily on regression-based model approaches, and there is insufficient guidance on how to appropriately apply its checklist items to machine learning prediction models. The issues we have identified regarding non-standardisation of reporting in deep learning research are increasingly becoming recognised as a barrier to robust evaluation of AI-based models. A step in the right direction was the Delphi process undertaken by Luo and colleagues10 to generate guidelines for developing and reporting machine learning predictive models. However, these guidelines have not been widely adopted, nor are they currently mandated by journals. An initiative to develop a machine learning version of the TRIPOD statement (TRIPOD-ML) was announced in April, 2019.10

Although most of the issues we have highlighted are avoidable with robust design and high-quality reporting, there are several challenges that arise in evaluating deep learning models that are specific to this field. The scale of data required for deep learning is a well recognised challenge. What is perhaps less recognised is the way that this requirement skews the types of data sources used in AI studies, and the relative paucity of some of the associated data. For example, in many studies, historical registry data collected from routine clinical care or open-source databases are used to supply sufficient input data. These image repositories are rarely quality controlled for the images or their accompanying labels, rendering the deep learning model vulnerable to mistakes and unidentified biases. Population characteristics for these large datasets are often not available (either due to not being collected, or due to issues of accessibility), limiting the inferences that can be made regarding generalisability to other populations and introducing the possibility of bias towards particular demographics.

Traditionally, heavy emphasis for developing and validating predictive models is on reporting all covariates and model-building procedures, to ensure transparent and reproducible, clinically useful tools.10 There are two main reasons why this is not possible in deep learning models in medical imaging. First, given the high dimensionality of the images, there are often too many individual datapoints driving predictions to identify specific covariates. Second, this level of influence and transparency of the algorithm is fundamentally incompatible with the black box nature of deep learning, where the algorithm’s decisions cannot be inspected or explained. Few methods for seeing inside the black box—the black box deconvolution—are available, but new methods are being actively explored. An important example is the use of saliency or heat maps, which many studies adopt to provide some qualitative assessment of predictive features within the image.42,110,111,112,113 Other recent approaches such as influence functions and segmentation can offer additional information alongside saliency or heat maps.42,110 There are two main reasons why this is not possible in deep learning models in medical imaging. First, given the high dimensionality of the images, there are often too many individual datapoints driving predictions to identify specific covariates. Second, this level of influence and transparency of the algorithm is fundamentally incompatible with the black box nature of deep learning, where the algorithm’s decisions cannot be inspected or explained. Few methods for seeing inside the black box—the black box deconvolution—are available, but new methods are being actively explored. An important example is the use of saliency or heat maps, which many studies adopt to provide some qualitative assessment of predictive features within the image.42,110,111,112,113 Other recent approaches such as influence functions and segmentation can offer additional information alongside saliency or heat maps.42,110
inability to interrogate a deep learning model, some caution should be exercised when making assumptions on a model’s generalisability. For example, an algorithm could incorrectly form associations with confounding non-pathological features in an image (such as imaging device, acquisition protocol, or hospital label) simply due to differences in disease prevalence in relation to those parameters.\textsuperscript{7,10,12} Another consideration is the transparency of reporting deep learning model building procedures. These studies often do not report the full set of hyperparameters used, meaning the model cannot be reproduced by others. There are also issues of underlying infrastructure that pose similar challenges. For example, those building the AI model might use custom-built or expensive infrastructure that is simply not available to most research groups, and thus present concerns around reproducibility and the ability to scrutinise claims made in peer review. Cloud-based development environments can support code sharing between researchers without compromising proprietary information, but more work is needed to establish gold standards in reporting results in this domain.

Any diagnostic test should be evaluated in the context of its intended clinical pathway. This is especially important with algorithms where the model procedures and covariates cannot be presented explicitly. A randomised head-to-head comparison to an alternative diagnostic test, in the context of a clinical trial, could reveal and quantify possible clinical implications of implementing an algorithm in real life. Moreover, a common problem of test evaluation research could be overcome by testing these algorithms within a clinical context: classification tasks are typically assessed in isolation of other clinical information that is commonly available in the diagnostic work-up.\textsuperscript{8} Prospective evaluations of diagnostic tests as complex interventions would not only reveal the impact of these algorithms upon diagnostic yield but also on therapeutic yield.\textsuperscript{7,8} In this context, the reporting of AI and machine learning interventional trials warrant additional consideration, such as how the algorithm is implemented and its downstream effects on the clinical pathway. In anticipation of prospective trials being the next step, extensions to the CONSORT and SPIRIT reporting guidelines for clinical trials involving AI interventions are under development.\textsuperscript{9,10,11,12}

Diagnosis of disease using deep learning algorithms holds enormous potential. From this exploratory meta-analysis, we cautiously state that the accuracy of deep learning algorithms is equivalent to health-care professionals, while acknowledging that more studies considering the integration of such algorithms in real-world settings are needed. The more important finding around methodology and reporting means the credibility and path to impact of such diagnostic algorithms might be undermined by an excessive claim from a poorly designed or inadequately reported study. In this review, we have highlighted key issues of design and reporting that investigators should consider. These issues are pertinent for ensuring studies of deep learning diagnostics—or any other form of machine learning—are of sufficient quality to evaluate the performance of these algorithms in a way that can benefit patients and health systems in clinical practice.

Contributors
AKD, ETF, JRL, KB, LF, LMB, MKS, PAK, and XL contributed to the conception and design of the study. AK, AB, CK, DJF, GM, LF, MS, TM, SKW, and XL contributed to the literature search and data extraction. AKD, ETF, JRL, KB, LF, LMB, MKS, PAK, and XL contributed to data analysis and interpretation. AK, AB, CK, DJF, ETF, GM, MS, and SKW contributed to critical revision of the manuscript. AKD, JRL, LF, LMB, TM, SKW, PAK, and XL contributed to writing the manuscript, and all authors approved the manuscript. AKD, LMB, and PAK guarantee the integrity of the work. LF and XL contributed equally to this work.

Declaration of interests
PAK is an external consultant for DeepMind. JRL is an employee of DeepMind Technologies, a subsidiary of Alphabet. ETF has received personal fees from Verily and Voxel Cloud. and declare no support from any organization for the submitted work. LF, XL, AK, SKW, DJF, TM, AB, BM, MS, CK, MKS, KB, LMB and AKD have nothing to disclose.

Data sharing
The search strategy and extracted data contributing to the meta-analysis is available in the appendix; any additional data are available on request.

Acknowledgments
PAK is supported by a National Institute for Health Research (NIHR) Clinician Scientist Award (NIHR-CS–2014–14–023). XL and AKD are supported by a Wellcome Trust Health Innovation Challenge Award (20014/Z/15/Z). The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health.

References
1. Fletcher KH. Matter with a mind; a neurological research robot. Research 1951; 4: 105–07.
2. Shoham Y, Perrault R, Brynjolfsson E, et al. The AI Index 2018 annual report. AI Index Steering Committee, Human-Centered AI Initiative, Standford, CA: Stanford University, 2018.
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017; 60: 84–90.
4. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42: 60–88.
5. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans Pattern Anal Mach Intell 2017; 39: 652–63.
6. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 2012; 29: 82–97.
7. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res 2011; 12: 2493–537.
8. Hadsell R, Erkan A, Sermanet P, Soffiater M, Muller U, LeCun Y. Deep belief net learning in a long-range vision system for autonomous off-road driving. 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems; Nice, France; Sept 22–26, 2008: 628–33.
9. Hadsell R, Sermanet P, Ben J, et al. Learning long-range vision for autonomous off-road driving. J Field Rob 2009; 26: 120–44.
10. Jia S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA 2016; 316: 2353–54.
11. Darcy AM, Louie AK, Roberts L. Machine learning and the profession of medicine. JAMA 2016; 315: 551–52.
12. Coiera E. The fate of medicine in the time of AI. Lancet 2018; 392: 2331–32.
13. Zhang L, Wang H, Li Q, Zhao M-H, Zhan Q-M. Big data and medical research in China. BMJ 2018; 360: j5910.
14. Schlemmer H-P, Bittencourt LK, D’Anastasi M, et al. Global challenges for cancer imaging. J Glob Oncol 2018; 4: 1–10.
Articles

15 King BF Jr. Artificial intelligence and radiology: what will the future hold? J Am Coll Radiol 2018; 15: 501-03.

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

17 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Syst 2010; 8: 1–9.

18 Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014; 11: e1001744.

19 Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019; 25: 954–61.

20 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. Br J Radiol 2018; 91: 20170726.

21 Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol 2018; 136: 803–10.

22 Schlegl T, Waldstein SM, Bogunovic H, et al. Fully-automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology 2018; 125: 549–58.

23 Harbord RM, Whiting P, Sterne JAC. et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008; 61: 1095–103.

24 Abbasi-Sureshjani S, Dashtihozorg B, ter Haar Romery BM, Fleuret F. Exploratory study on direct prediction of diabetes using deep residual networks. In: VipIMAGE 2017. Basel: Springer International Publishing, 2018: 797–802.

25 Adams M, Chen W, Huldofd M, McCusker MW, Howe PD, Gaillard F. Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 2019; 63: 27–32.

26 Ariji Y, Fukuda M, Kise Y, et al. Contract-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 Surg Oral Med Oral Pathol Oral Radiol 2019; 127: 458–63.

27 Ayed NGB, Masnoudi AD, Sellami D, Abid R. New developments in the diagnostic procedures to reduce prospective biopsies breast. 2015 International Conference on Advances in Biomedical Engineering (ICABME). Beirut, Lebanon; Sept 16–18, 2015: 205–08.

28 Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frausenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multicentre image analysis software in the detection of breast cancer. Invest Radiol 2017; 52: 434–40.

29 Biet N, Raiipurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018; 15: e1002699.

30 Brinker TJ, Helekzer 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. Eur J Cancer 2019; 111: 148–54.

31 Burdini P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD severity using universal deep features and transfer learning for automated AMD analysis. Comput Biol Med 2017; 82: 80–86.

32 Burdini P, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Utility of deep learning methods for referability classification of age-related macular degeneration. JAMA Ophthalmol 2018; 136: 1105–07.

33 Burdini 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 Ophthalmol 2018; 136: 1105–07.

34 Byra M, Galperin M, Oyeda-Fournier H, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys 2019; 46: 746–55.

35 Cao R, Baigiran AM, Mirak SA, et al. Joint prostate cancer detection and gleason score prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging 2019; published online Feb 22. DOI:10.1109/TMI.2019.2901928.

36 Chee CG, Lim 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 Am J Roentgenol 2019; published online March 27. DOI:10.2214/AJR.18.20817.

37 Choi JS, Han B-K, 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 J Radiol 2019; 20: 749.

38 Choi RJ, 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: 688–97.

39 Cioni P, Chung K, van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep 2017; 7, 46879.

40 Codella NCF, Nguyen QB, Pankanti S, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J Res Dev 2017; 61: 5–15.

41 Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histology images using deep learning. Nat Med 2018; 24: 1559–67.

42 De Fauw J, Lednam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018; 24: 1342–50.

43 Ding Y, Sohn JH, Kawczynski MG, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology 2019; 290: 456–64.

44 Dunning JA, Yi D, Langlois CP, Re C, Rubin DL, Langren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2019; 290: 537–44.

45 Elsheshai 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: 2199–210.

46 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–18.

47 Fujioka T, Kuhota K, Mori M, et al. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. Jpn J Radiol 2019; 37: 466–72.

48 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. Br J Dermatol 2019; 180: 373–81.

49 Gómez-Valverde JJ, Antón A, Fatti G, et al. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. Biomed Opt Express 2019; 10: 892–911.

50 Grewal M, Srivastava MM, Kumar P, Varadarajan S. RADnet: radiologist-level accuracy using deep learning for hemorrhage detection in CT scans. 2018 IEEE 13th International Symposium on Biomedical Imaging (ISBI 2018). Washington, DC, USA; April 4–7, 2018: 281–84.

51 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. Ann Oncol 2018; 29: 1836–42.

52 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. Eur Radiol 2019; 29: 3338–42.

53 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. J Invest Dermatol 2018; 138: 1529–38.

54 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: e0194193.

55 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. Clin Infect Dis 2018; published online Nov 12. DOI:10.1093/cid/ciy962.
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.

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: e191995.

Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018; 172: 3112–31.e9.

Kim SM, Han H, Park JM, et al. A comparison of logistic regression analysis and an artificial neural network using the BI-RADS lexicon for ultrasonography in conjunction with introserver variability. J Digit Imaging 2012; 25: 599–606.

Kim K, Kim S, Lee YH, Lee SH, Lee HS, Kim S. Performance of the deep convolutional neural network based magnetic resonance image scoring algorithm for differentiating between tuberculoid and pyogenic spondylitis. Sci Rep 2018; 8: 13124.

Kim Y, Lee KJ, Sunwoo L, et al. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest Radiol 2019; 54: 7–15.

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. Dentomaxillofac Radiol 2019; published online May 22. DOI:10.1259/dfr.2019009.

Ko SY, Lee JH, Yoon JH, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. Head Neck 2019; 41: 885–91.

Kumagai Y, Takubo K, Kawada K, et al. Diagnosis using deep-learning artificial intelligence based on the endoscopic observation of the esophagus. Esophagus 2019; 16: 180–87.

Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from CT images. NeuroImage 2019; 193–201.

Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. Cancer Commun 2018; 38: 59.

Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep learning. World J Gastroenterol 2019; 25: 193–201.

Lin CM, Hou YL, Chen TY, Chen KH. Breast nodules computer-aided diagnostic system design using fuzzy cerebral model neural networks. IEEE Trans Fuzzy Syst 2014; 22: 693–99.

Lindsey R, Daluiski A, Chopra S, et al. Deep neural network diagnostic system design using fuzzy cerebellar model neural networks. IEEE J Biomed Health Inform 2019; 58–65.

Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat Biomed Eng 2017; 1: 0024.

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. Transl Vis Sci Technol 2018; 7: 41.

Matsuba S, Tabuchi H, Obusgi H, et al. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. Int Ophthalmol 2019; 39: 1269–75.

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. Gastrointest Endosc 2019; published online May 8. DOI:10.1016/j.gie.2019.04.245.

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: 218–28.

Olcakz F, Fuhlerberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthopaedica 2017; 88: S81–S86.

Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. Ophthalmology 2019; 126: 565–75.

Poedjioatnoet W, Suebnukarn S. Application of convolutional neural network in the diagnosis of jaw tumors. Healthc Inform Res 2018; 24: 231–36.

Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the ChestX-Net algorithm to practicing radiologists. PLoS Med 2018; 15: e1002686.

Ruanvivinsook P, Krause J, Chotcomwong P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. NPJ Digit Med 2019; 2: 25.

Sayres R, Taly A, Rahimi E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 2019; 126: 552–64.

Shibutani T, Nakajima T, Wakahayashi H, et al. Accuracy of an artificial neural network for detecting a regional abnormality in myocardial perfusion SPECT. Ann Nucl Med 2019; 33: 86–92.

Shichijo S, Nomura S, Aoyama K, et al. Application of convolutional neural networks in the diagnosis of Helicobacter pylori infection based on endoscopic images. EobioMedicine 2017; 25: 106–11.

Singh R, Kalra MK, Sittwangkaj C, et al. Deep learning in chest radiography: detection of findings and presence of change. PLoS One 2018; 13: e0204155.

Song W, Li S, Liu J, et al. Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE Biomed Health Inform 2019; 23: 1215–24.

Stoffel E, Becker AS, Warming MC, et al. Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis. Eur J Radiol Open 2018; 5: 163–70.

Streba CT, Ionescu M, Gheonea DI, et al. Contrast-enhanced ultrasonography parameters in neural network diagnosis of liver tumors. World J Gastroenterol 2012; 18: 4827–34.

Sun L, Li Y, Zhang YT, et al. A computer-aided diagnostic algorithm improves the accuracy of transesophageal echocardiography for left atrial thrombi: a single-center prospective study. J Ultrasound Med 2014; 33: 83–91.

Tsichandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019; 155: 58–65.

Uraoka 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 Radiol 2019; 48: 239–44.

van Geemen MJJP, van Ginkel 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 Trans Med Imaging 2016; 35: 1273–84.

Walsh SIF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease at high-resolution computed tomography: a case-colort study. Lancet Respir Med 2018; 6: 837–45.

Wang HK, Zhou ZW, Li YC, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from F-18-FDG PET/CT images. EJNMMI Res 2017; 7: 11.

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. Quant Imaging Med Surg 2018; 8: 491–99.

Wang L, Yang S, Yang S, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J Surg Oncol 2019; 17: 12.

Wright JW, Duguid R, McKeeie F, Staff RT. Automatic classification of DMSA scans using an artificial neural network. Phys Med Biol 2014; 59: 779–800.

Wu L, Zhou W, Han X, et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. Endoscopy 2019; 51: 322–31.

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. Eur Radiol 2019; published online April 30. DOI:10.1007/s00330-019-06163-2.

Yu C, Yang S, Kim W, et al. Acrail melanoma detection using a convolutional neural network for dermoscopy images. PLoS One 2018; 13: e0193321.
Articles

99 Zhang C, Sun X, Dang K, et al. Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. *Oncologist* 2019; published online April 17. DOI:10.1634/theoncologist.2018–0908.

100 Zhang Y, Wang L, Wu Z, et al. Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images. *IEEE Access* 2019; 7: 10232–41.

101 Zhao W, Yang J, Sun Y, et al. 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer Res* 2018; 78: 6881–89.

102 Riet GT, Ter Riet G, Bachmann LM, Kessels AGH, Khan KS. Individual patient data meta-analysis of diagnostic studies: opportunities and challenges. *Evid Based Med* 2013; 18: 165–69.

103 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; 19: 453–73.

104 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; 18: e323.

105 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019; 393: 1577–79.

106 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; 35: 1925–31.

107 Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; 3: 141–52.

108 Koh PW, Liang P. Understanding black-box predictions via influence functions. *Proc Mach Learn Res* 2017; 70: 1885–94.

109 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018; 15: e1002683.

110 Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019; 2: 31.

111 Bachmann LM, ter Riet G, Weber WE, Kessels AGH. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. *J Clin Epidemiol* 2009; 62: 357–361.e2.

112 Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012; 344: e686.

113 EQUATOR Network. Reporting guidelines under development. http://www.equator-network.org/library/reporting-guidelines-under-development (accessed Aug 4, 2019).

114 Liu X, Faes, L, Calvert MJ, Denniston AK, CONSORT-AI/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. *Lancet* (in press).

115 The CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* (in press).