What are the Quality Assessment Standards used in Artificial Intelligence Diagnostic Accuracy Systematic Reviews?

Shruti Jayakumar
Department of Surgery and Cancer, Imperial College London, United Kingdom
Viknesh Sounderajah
Department of Surgery and Cancer, Imperial College London, United Kingdom
Pasha Normahani
Department of Surgery and Cancer, Imperial College London, United Kingdom
Leanne Harling
Department of Surgery and Cancer, Imperial College London, United Kingdom
Sheraz R. Markar
Department of Surgery and Cancer, Imperial College London, United Kingdom
Hutan Ashrafian (hutan@researchtrials.net)
Department of Surgery and Cancer, Imperial College London, United Kingdom
Ara Darzi
Department of Surgery and Cancer, Imperial College London, United Kingdom

Research Article

Keywords: Diagnostic Accuracy, Artificial Intelligence, Quality Assessment

Posted Date: March 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-329433/v1

What are the Quality Assessment Standards used in Artificial Intelligence Diagnostic Accuracy Systematic Reviews?

Shruti Jayakumar[1,2], Viknesh Sounderajah[1,2], Pasha Normahani[1,2], Leanne Harling[1,3], Sheraz R. Markar[1,2], Hutan Ashrafian[1,2,3], Ara Darzi[1,2]

[1]Institute of Global Health Innovation, Imperial College London, United Kingdom
[2]Department of Surgery and Cancer, Imperial College London, United Kingdom
[3]Department of Thoracic Surgery, Guy's Hospital, London, United Kingdom

[3]Corresponding Author
Mr Hutan Ashrafian MRCS PhD MBA
Institute of Global Health Innovation
Imperial College London
London SW7 2AZ
Email: hutan@researchtrials.net

Word Count: 5422 words (7630 words including references)

Figures and Tables: 5 figures, 8 tables

Key Words: Diagnostic Accuracy, Artificial Intelligence, Quality Assessment

Introduction

With ever expanding applications for the use of artificial intelligence (AI) in healthcare, interest in its capabilities to analyse and interpret diagnostic tests has increased. AI driven approaches to interpretation of diagnostic tests has potential to overcome several current limitations on clinician availability, time to diagnosis and diagnostic accuracy. Recently, various deep learning algorithms have been shown to mimic human-like or demonstrate superhuman performance in analysis of radiological findings.[1] In conjunction with AI, radiologists are capable of improving sensitivity and specificity as well as minimising inter- and intra-observer variability in interpretation. Similar studies have also been in conducted in non-radiological diagnostics, including AI-driven analysis of endoscopic, retinal and histopathological images.[2–4] As studies examining AI driven approaches to diagnostic interpretation have become prevalent, systematic reviews have increasingly been published to amalgamate and report these results. Given the diversity and heterogeneity of existing AI techniques, with further rapid expansion expected, clinicians and policy makers may find it difficult to interpret these results and implement these models in their clinical practice. Additionally, it is prudent to ensure included studies are of high methodological quality and employ rigorous standards of outcome reporting, as they may be influential in altering guidelines or prompting significant policy change. On the other hand, poor quality studies with a lack of transparent reporting may lead to scepticism within healthcare professionals and members of the public therefore leading to unnecessary delays in technological adoption. It is therefore imperative that authors of systematic reviews critically appraise literature using an evidence-based, validated quality assessment tool to enable adequate comparison between studies.

The most commonly used tool for the methodological assessment of secondary research studies remains the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool. QUADAS was developed in 2003 and updated in 2011. The original tool comprised of fourteen items on patient selection and spectrum, reference standard, presence of various biases, test execution, study withdrawals and indeterminate results.[5] The updated version was modified to categorise the questions into four key domains: (i) patient selection, (ii) index test, (iii) reference standard and (iv) flow and timing, with each domain assessed for biases and the first three also assessed for applicability.[6] However, the applicability of QUADAS for AI specific studies is unknown. These studies differ methodologically from conventional trials and consist of distinctive features, techniques and a different entity of analytical challenges. Given the differences in study design and outcome reporting, the areas of potential bias are also likely to differ substantially. However, despite these assumptions, there have been no formal studies examining the adherence and suitability of QUADAS in this genre of studies. Moreover, there has not been a similar evaluation with respect to emerging AI centred quality appraisal tools, such as the Radiomics Quality Score (RQS), which was specifically designed for studies reporting on algorithm-based extraction of features from medical images.[7]

Therefore, the primary aim of this meta-research study is to evaluate adherence of QUADAS within systematic reviews of AI-based diagnostic accuracy. The secondary aims include (i) assessing the applicability of QUADAS for AI-based diagnostic accuracy studies, (ii) identifying other tools for methodological quality assessment and (iii) identifying key features that an AI specific quality assessment tool should incorporate.

Materials and Methods

Search strategy

An electronic search was conducted for studies in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines to identify systematic reviews reporting on diagnostic accuracy studies in AI studies (Figure 1).[8] MEDLINE was searched from January 2000 – December 2020 using a mixture of keywords and MeSH terms. The search strategy consisted of systematic reviews related to artificial intelligence (artificial intelligence, machine learning and deep learning) and diagnostic accuracy.

Study selection

Two independent reviewers screened titles and abstracts for initial inclusion. Studies were included if they meet the following inclusion criteria: (1) systematic review (2) reporting on diagnostic accuracy in AI studies. Commentary articles, conference extracts and narrative reviews were excluded. Studies either examining prognostication or reporting on AI/ML to predict the presence of disease were also excluded. Two reviewers (SJ and VS) independently screened titles and abstracts for potential inclusion. All potential abstracts were subjected to full-text review by two independent reviewers. Disagreements were resolved through discussion with a third independent reviewer (HA).

Data extraction

Data was extracted onto a standardised proforma by two independent reviewers (VS and SJ). Study characteristics extracted were study author, year, institution, country, journal and journal impact factor. Data was collected on use of QUADAS and/or other quality assessment tools, quality assessment tool adherence, modifications to pre-existing tools, use of multiple tools to improve applicability to AI specific studies and any limitations pertaining to quality assessment expressed by study authors.

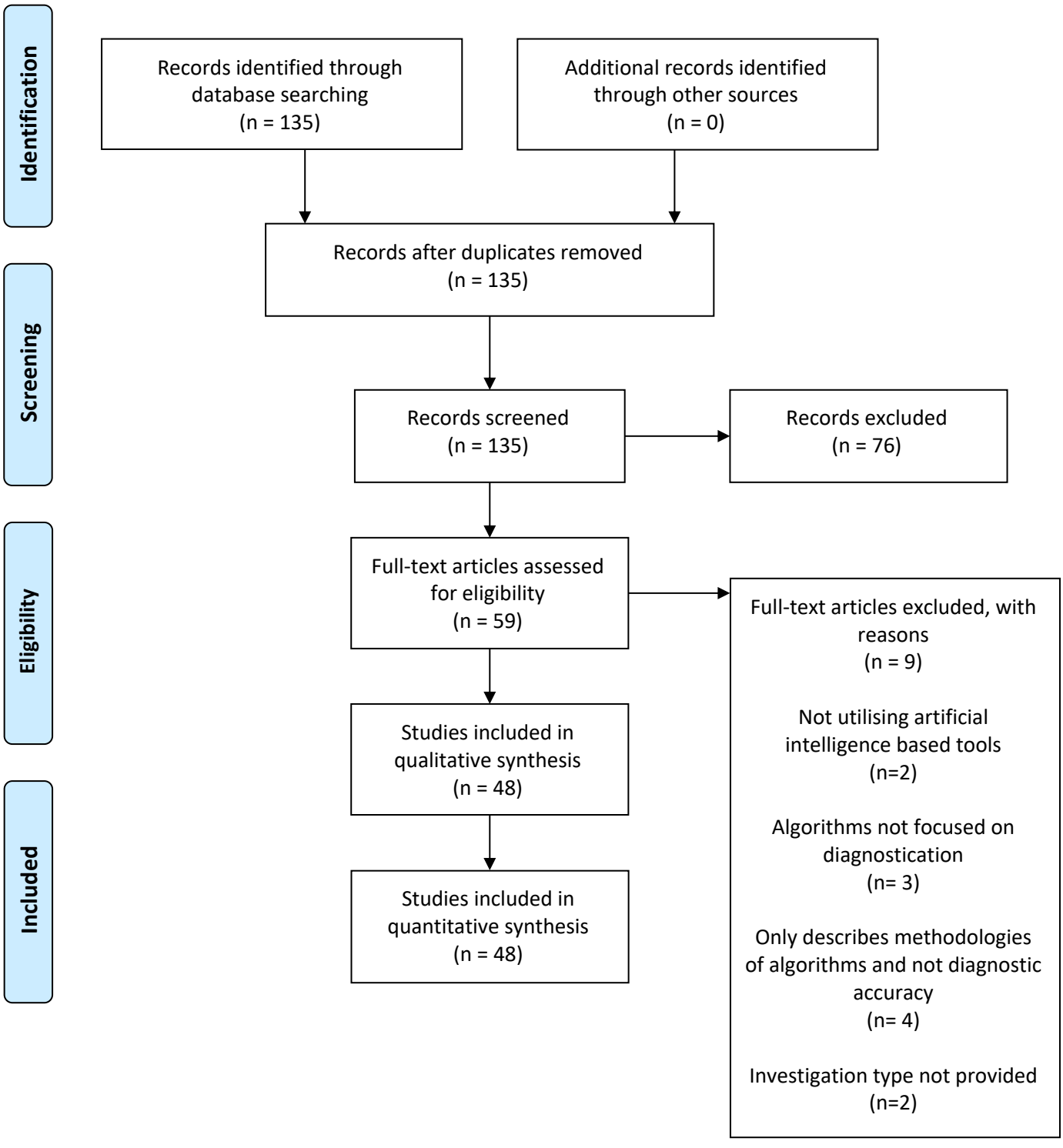
Studies were classified into four clinical categories based upon the type of sample evaluated and upon the diagnostic task: (a) axial medical imaging, (b) non-axial medical imaging, (c) histopathological digital records (digital pathology) and (d) photographic images.

Key AI-related extraction items were identified through examination of recently developed AI extensions to existing quality assessment tools. A consensus was reached amongst authors to ascertain vital items.

Quality Assessment

The AMSTAR 2 (A MeaSurement Tool to Assess systematic Reviews) was employed to evaluate the quality of included studies (Supplementary Table 1).[9]

Figure 1: PRISMA Guidelines



Results

The search yielded 135 papers, of which 49 met the eligibility criteria (Figure 1). Three papers were excluded upon full-text review as the systematic reviews focussed upon prediction models. Two papers were excluded due to a lack of focus on artificial intelligence-based diagnostics. Four studies were excluded as they solely discussed the types and methodologies of AI based tools. One study was excluded as it did not specify imaging type.

Study Characteristics

A total of 1110 studies were included across all 48 systematic reviews, with an average of 23 studies within each systematic review (range: 2 – 111 studies). The full study characteristics are provided in Tables 1 to 4. Twenty-three reviews analysed axial imaging, nine analysed non-axial imaging, three analysed digital pathology , two analysed electrocardiograms and fifteen analysed photographic images. Of these photographic images, six analysed endoscopic images, four analysed skin lesions and five analysed fundus photography or optical coherence tomography.

The most common artificial intelligence techniques used within the studies comprising the systematic reviews include support vector machines and artificial neural networks, specifically convolutional neural networks.

Quality Assessment

36 reviews (75% of studies) undertook a form of quality assessment, of which 27 utilised the QUADAS-2 tool. Further breakdown of quality assessment by study category is detailed below.

Diagnostic Accuracy of AI in Axial Imaging

23 systematic reviews comprising 621 studies reported on the application of AI models to diagnostic axial imaging (Table 1). Of the 23 studies, 14 performed quality assessment with 7 reporting use of the QUADAS tool (Table 5). One study utilised RQS and another study utilised the RQS in addition to QUADAS. Other quality assessment tools used include MINORS (n=3), the Newcastle-Ottawa Score (n=2) and the Jadad Score (n=2).

Out of the seven studies employing QUADAS, five studies completely reported risk of bias and applicability as per the QUADAS guidelines whilst one study only reported on risk of bias. One study provided QUADAS ratings given by each of the study authors; but did not provide a consensus table.[10]

Four studies modified the existing quality assessment tools to improve suitability and applicability of the tool. Cho et al. tailored the QUADAS tool by applying select signalling questions from CLAIM (Checklist for Artificial Intelligence in Medical Imaging).[11] Pellegrini and colleagues reported difficulties in finding a suitable quality assessment tool for machine

learning diagnostic accuracy reviews and selectively applied items in the QUADAS tool to widen study inclusion.[12] One study modified the MINORS checklist whilst another study used a modified version of the MINORS checklist in addition to TRIPOD.[13,14]

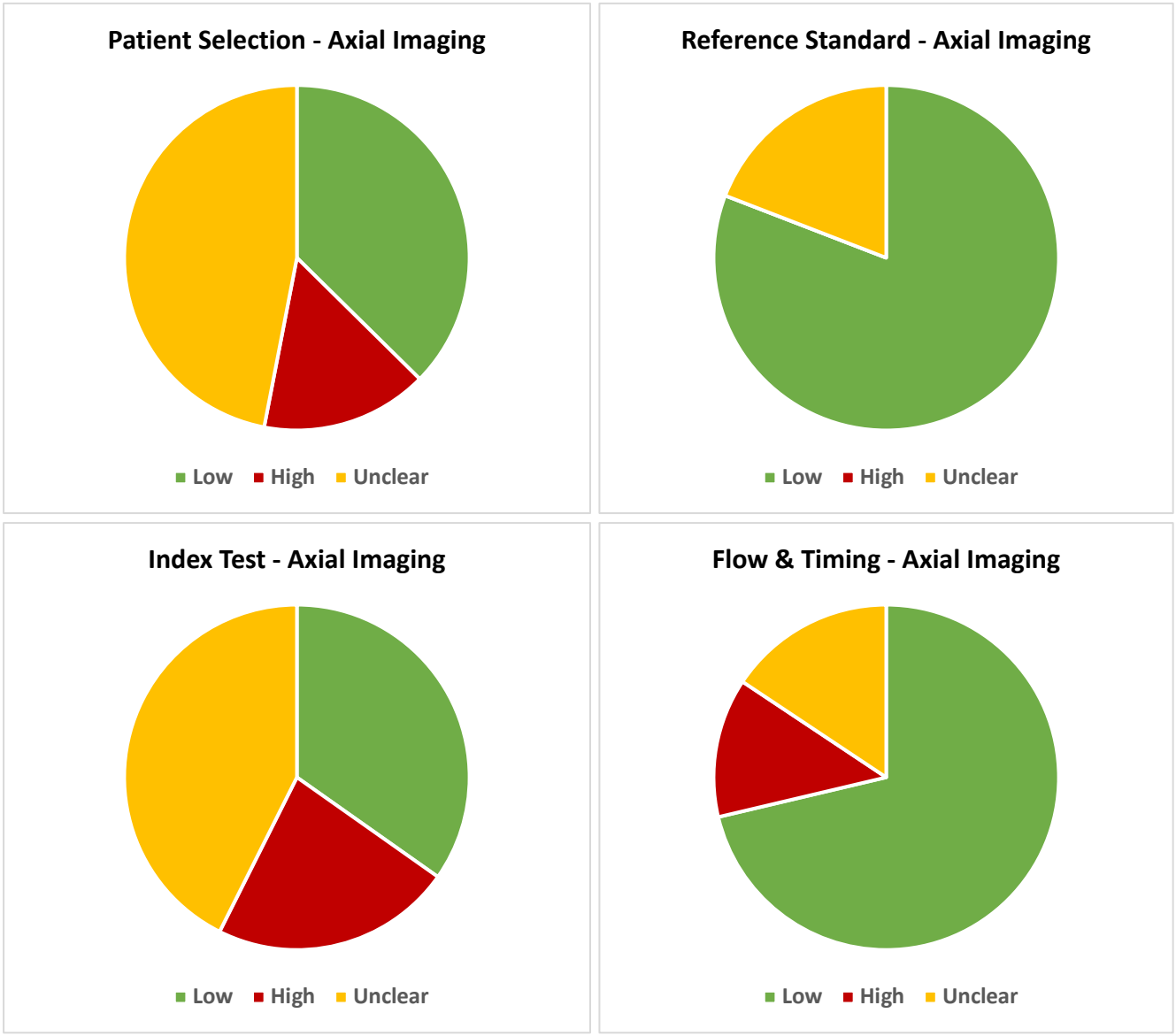


Figure 2: Pie charts demonstrating the risk of bias amongst axial imaging studies, as assessed through QUADAS

Table 1: Systematic Reviews of Artificial Intelligence Based Diagnostic Accuracy Studies in Axial Imaging

Author	Specialty	Included Studies	Input Variables	Diagnosis
Nayantara 2020[15]	Hepatology	25	CT	Liver lesions
Cho 2020[11]	Oncology	12	MRI	Cerebral metastases
Crombé 2020[16]	Oncology	52	CT, CT-PET, MRI, US	Sarcoma
Kunze 2020[17]	Musculoskeletal	11	MRI	ACL and/or meniscal tears
Groot 2020[14]	Musculoskeletal	14	MRI, X-Rays, US	X-Ray: Fracture detection and/or classification MRI: meniscal/ligament tears, tuberculous vs pyogenic spondylitis US: lateral epicondylitis
Steardo Jr 2020[24]	Psychiatry	22	fMRI	Schizophrenia
Ninatti 2020[29]	Oncology	24	CT, PET-CT	Molecular therapy targets
Ursprung 2020[10]	Oncology	57	CT, MRI	Renal cell carcinoma
Halder 2020[36]	Respiratory Medicine	45	CT	Lung nodules
Li 2019[39]	Respiratory Medicine	26	CT	Lung nodule detection and/or classification
Azer 2019[38]	Hepatology / Oncology	11	CT, MRI, US, Pathology slides	Hepatocellular carcinoma, liver masses
Jo 2019[41]	Neurology	16	MRI, PET, CSF	Alzheimer's disease
Moon 2019[40]	Psychiatry	43	sMRI, fMRI	Autism spectrum disorder
Sarmento 2020[43]	Neurology	8	CT or MRI	Stroke
Filippis 2019[45]	Psychiatry	35	sMRI, fMRI	Schizophrenia
Langerhuizen 2019[13]	Musculoskeletal	10	CT, X-Rays	Fracture detection and/or classification
Pellegrini 2018[12]	Neurology	111	MRI, CT	Mild cognitive impairment, dementia
Pehrson 2019[47]	Respiratory Medicine	19	CT	Lung nodule
Bruin 2019[54]	Psychiatry	12	sMRI, fMRI	Obsessive compulsive disorder
McCarthy 2018[52]	Neurology	28	MRI	Frontotemporal dementia
Nguyen 2018[51]	Neurology / Oncology	8	MRI	Differentiate glioblastoma and primary CNS lymphoma
Senders 2018[55]	Neurosurgery	14	CT, MRI, history, age, gender	Intracranial masses, tumours
Smith 2017[56]	Musculoskeletal	18	sMRI, fMRI	Musculoskeletal pain

Table 2: Systematic Reviews of Artificial Intelligence Based Diagnostic Accuracy Studies in Non-Axial Imaging

Author	Specialty	Included Studies	Input Variables	Diagnosis
Li 2020[21]	Respiratory Medicine	15	Chest X-Ray	Pneumonia
Xu 2020[20]	Oncology / Endocrinology	19	US	Malignant thyroid nodules
Yang 2020[26]	Musculoskeletal	9	X-Rays	Fractures
Groot 2020[14]	Musculoskeletal	14	MRI, X-Rays, US	X-Ray: Fracture detection and/or classification MRI: meniscal/ligament tears, tuberculous vs pyogenic spondylitis US: lateral epicondylitis
Li 2020[30]	Oncology	10	US	Malignant breast masses
Azer 2019[38]	Hepatology / Oncology	11	CT, MRI, US, Pathology slides	Hepatocellular carcinoma, liver masses
Harris 2019[42]	Respiratory Medicine	53	Chest X-Ray	Tuberculosis
Zhao 2019[44]	Endocrinology	5	Ultrasound	Thyroid nodules
Langerhuizen 2019[13]	Musculoskeletal	10	X-Rays, CT	Fracture detection and/or classification

Table 3: Systematic Reviews of Artificial Intelligence Based Diagnostic Accuracy Studies in Photographic Images

Author	Specialty	Included Studies	Input Variables	Diagnosis
Bang 2020[18]	Gastroenterology	8	Endoscopic images	H. Pylori infection
Mohan 2020[22]	Gastroenterology	9	Endoscopic images	Gastrointestinal ulcers/haemorrhage
Hassan 2020[25]	Gastroenterology	5	Colonoscopic images	Polyps
Lui 2020[34]	Gastroenterology	18	Colonoscopy images	Polyps
Lui 2020[27]	Gastroenterology	23	Endoscopic images	Neoplastic lesions, Barrett's oesophagus, squamous oesophagus, H. Pylori status
Wang 2020[28]	Ophthalmology	24	Fundus photography	Diabetic Retinopathy
Soffer 2020[31]	Gastroenterology	10	Wireless Capsule Endoscopic images	Detection of ulcers, polyps, bleeding, angioectasia
Islam 2020[32]	Ophthalmology	31	Fundus photography	Retinal vessel segmentation
Islam 2020[35]	Ophthalmology	23	Fundus photography	Diabetic retinopathy
Murtagh 2020[37]	Ophthalmology	23	OCT, Fundus photography	Glaucoma
Nielsen 2019[46]	Ophthalmology	11	Fundus photography	Diabetic Retinopathy
Marka 2019[48]	Dermatology / Oncology	39	Images of skin lesions	Non-melanoma skin cancer
Ruffano 2018[49]	Dermatology / Oncology	42	Images of skin lesions	Non-melanoma skin cancer
Chuchu 2018[50]	Dermatology / Oncology	2	Images of skin lesions	Melanoma
Rajpara 2009[58]	Dermatology / Oncology	30	Images of skin lesions	Melanoma

Table 4: Systematic Reviews of Artificial Intelligence Based Diagnostic Accuracy Studies in Pathology Images

Author	Specialty	Included Studies	Input Variables	Diagnosis
Azam 2020[19]	Pathology	25	Histology samples	Varied - dysplasia, malignancy, challenging diagnoses, identification of small objects, miscellaneous
Mahmood 2020[23]	Oncology / ENT/ Maxfax	11	Histology samples	Malignant head and neck lesions
Azer 2019[38]	Hepatology / Oncology	11	CT, MRI, US, Pathology slides	Hepatocellular carcinoma, liver masses

Table 5: Quality Assessment and Adherence to QUADAS in Systematic Reviews of Diagnostic Accuracy of Artificial Intelligence in Axial Imaging

Study	Input Variables	Quality Assessment Performed	QUADAS Used	Modifications to Existing Tools	Other Tools Used	QUADAS Results Reported
Nayantara 2020	CT	No	-	-	-	-
Halder 2020	CT	No	-	-	-	-
Azer 2019	CT, MRI, US, Pathology slides	No	-	-	-	-
Li 2019	CT	No	-	-	-	-
Jo 2019	MRI, PET, CSF	No	-	-	-	-
Sarmento 2019	CT, MRI	No	-	-	-	-
Pehrson 2019	CT	No	-	-	-	-
Bruin 2019	sMRI, fMRI	No	-	-	-	-
Senders 2018	CT, MRI History/age/gender	No	-	-	-	-
Langerhuizen 2019	X-Rays, CT	Yes	No	Yes – modified MINORS	MINORS	-
Smith 2017	sMRI, fMRI	Yes	No	No	Newcastle-Ottawa Scale	-
Crombe 2020	CT, MRI, US	Yes	No	No	Radiomics Quality Score	-
Kunze 2020	MRI	Yes	No	No	MINORS	-
Groot 2020	MRI, X-Rays, US	Yes	No	Yes – modified MINORS	MINORS, TRIPOD	-
Steardo Jr 2020	fMRI	Yes	No	No	Jadad	-
Filippis 2019	sMRI, fMRI	Yes	No	No	Jadad	-
Ninatti 2020	CT, PET-CT	Yes	Yes	No	TRIPOD	Yes
Cho 2020	MRI	Yes	Yes	Yes – modified QUADAS using CLAIM	CLAIM checklist for AI	Yes
McCarthy 2018	MRI	Yes	Yes	No	No	Yes
Moon 2019	sMRI, fMRI	Yes	Yes	No	No	Yes
Pellegrini 2018	MRI, CT	Yes	Yes	Yes – only used QUADAS criteria authors deemed applicable	No	Yes
Nguyen 2018	MRI	Yes	Yes	No	No	Yes (only for bias)
Ursprung 2019	CT, MRI	Yes	Yes	No	Radiomics Quality Score	Yes (multiple raters; no consensus)

Among the 115 studies across six systematic reviews, the patient selection was deemed to pose the highest or most unclear risk of bias. 54 of 115 studies (47%) were considered to have an unclear risk and 16 studies (14%) were classified as high risk of bias (Figure 2). A high proportion of studies were also considered to pose an unclear risk in the index test domain. 81% of studies had a low risk of bias in the reference standard domain with the remainder representing an unclear risk. Concern regarding applicability was generally low for most studies across all five reviews with 78.5%, 87.9% and 93.5% of studies having low concerns of applicability in the patient selection, index test and reference standard domains, respectively.

Diagnostic Accuracy of AI in Non-axial Imaging

9 systematic reviews comprising 146 studies reported on the application of AI models to non-axial imaging comprising X-Rays or Ultrasounds (Table 2). Three reviews additionally included studies that also reported on axial imaging.

Of the nine systematic reviews, seven performed quality assessment with five utilising QUADAS (Table 6). The remaining two studies utilised modified versions of the MINORS tools, with one of the studies also utilising TRIPOD as reported under Axial Imaging.

Table 6: **Quality Assessment and Adherence to QUADAS in Systematic Reviews of Diagnostic Accuracy of Artificial Intelligence in Non-Axial Imaging**

Study	Modality	Quality Assessment	QUADAS Used	Modifications to Existing Tools	Other Tools Used	QUADAS Results Reported
Li 2020	Chest X-Ray	No	-	-	-	-
Azer 2019	CT, MRI, US, Pathology slides	No	-	-	-	-
Langerhuizen 2019	X-Rays, CT	Yes	No	Yes	Modified MINORS	-
Groot 2020	MRI, X-Rays, US	Yes	No	Yes (modified MINORS)	TRIPOD + modified MINORS	-
Xu 2020	US	Yes	Yes	No	No	Yes
Yang 2020	X-Rays	Yes	Yes	No	No	Yes
Li 2020	US	Yes	Yes	No	No	Yes
Harris 2019	Chest X-Ray	Yes	Yes	No	No	Yes
Zhao 2019	US	Yes	Yes	No	No	Yes

Among the 89 studies across five systematic reviews, the index test domain posed the highest risk of bias whilst the patient selection domain posed the most unclear risk of bias (Figure 3). Concern regarding applicability was generally low for most studies across all five reviews with 79.1%, 79.1% and 90.7% of studies having low concerns of applicability in the patient selection, index test and reference standard domains, respectively.

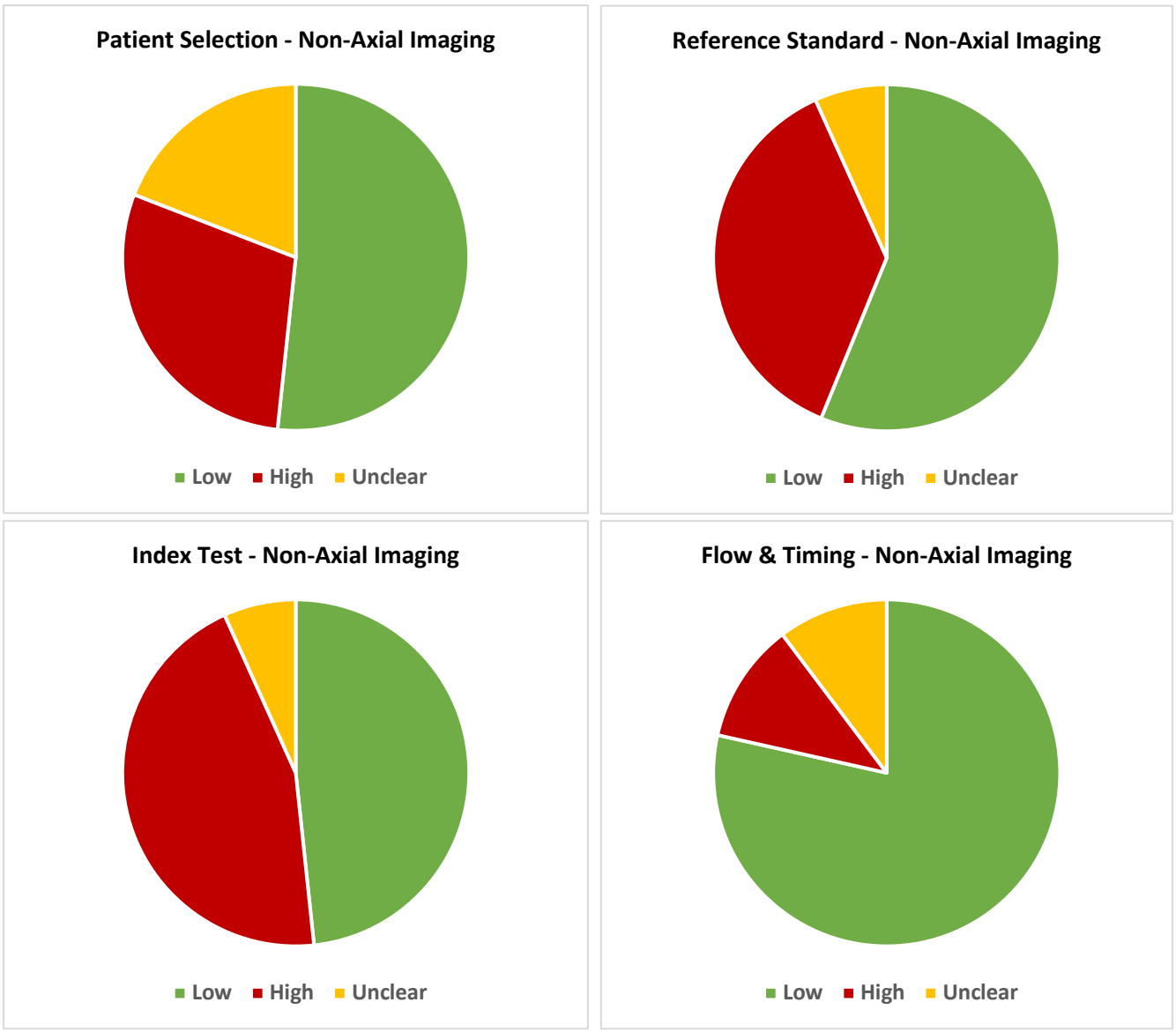


Figure 3: Pie charts demonstrating the risk of bias amongst non-axial imaging studies, as assessed through QUADAS

Diagnostic Accuracy of AI in Photographic Images

Fifteen systematic reviews comprising 316 studies reported on the application of AI to photo-based diagnostics (Table 3). This consisted of images of skin lesions (n=4), endoscopic images (n=6), and fundus photography or optical coherence tomography (n=5).

Of the fifteen systematic reviews, thirteen performed quality assessment with eleven utilising QUADAS (Table 7). One study did not report any details on QUADAS whilst another did not report on applicability concerns and only risk of bias. The remaining two studies utilised the Cochrane Risk of Bias Tool and modified version of the Newcastle-Ottawa scale. In addition, Ruffano et al. and Chuchu et al. adapted the QUADAS tool specifically for non-melanoma skin cancer and melanoma respectively with definitions and thresholds specified by consensus for low and high-risk for bias.

Among the 231 studies across 11 systematic reviews, the patient selection domain contained the highest risk of bias whilst the flow and timing domain posed the most unclear risk of bias (Figure 4). Concern regarding applicability was high or unclear in the patient selection domain for the majority of studies with 54.8% of studies reporting high or unclear applicability concerns. Concerns of applicability were lower in the index test and reference standard domain with 67.5% of studies reporting low concerns in the index test domain and 53.8% in the reference standard domain.

Table 7: Quality Assessment and Adherence to QUADAS in Systematic Reviews of Diagnostic Accuracy of Artificial Intelligence in Photographic Images

Study	Modality	Quality Assessment	QUADAS Used	Modifications to Existing Tools	Other Tools Used	QUADAS Results Reported
Mohan 2020	Endoscopic images	No	-	-	-	-
Rajpara 2009	Images of skin lesions	No	-	-	-	-
Hassan 2020	Real-time computer-aided detection colonoscopy	Yes	No	No	Cochrane Risk Bias Tool	-
Murtagh 2020	OCT/Fundus photophraphy	Yes	No	Yes - modified Newcastle-Ottawa Scale	Newcastle-Ottawa Scale	-
Bang 2020	Endoscopic images	Yes	Yes	No		Yes
Lui 2020	Endoscopic images	Yes	Yes	No		Yes
Wang 2020	Fundus photography	Yes	Yes	No		Yes
Soffer 2020	Wireless capsule endoscopy	Yes	Yes	No		Yes - but not for applicability
Islam 2020	Fundus photography	Yes	Yes	No		Yes
Lui 2020	Colonoscopy	Yes	Yes	No		Yes
Islam 2020	Fundus photography	Yes	Yes	No		Yes
Nielsen 2019	Fundus photography	Yes	Yes	No		Yes
Marka 2019	Images of skin lesions	Yes	Yes	No		Yes
Ruffano 2018	Images of skin lesions	Yes	Yes	Yes - modified for non-melanoma skin cancers		Yes
Chuchu 2018	Images of skin lesions	Yes	Yes	Yes - modified for melanoma		Yes

Diagnostic Accuracy of AI in Pathology

Three systematic reviews comprising 47 studies reported on the application of AI to pathology. One review examined pathology slides in addition to imaging (Table 4).

Two reviews performed quality assessment utilising QUADAS (Table 8). Mahmood et al a tailored QUADAS-2 tool. Only one review provided a tabular display of QUADAS assessment in the recommended format[19] and reported low risk of bias among the majority of included studies across all domains (Patient Selection: 64% of studies low risk; Index Test: 80% low risk;

Reference Standard: 92% low risk; Flow and Timing: 84% low risk) and low concerns regarding applicability.

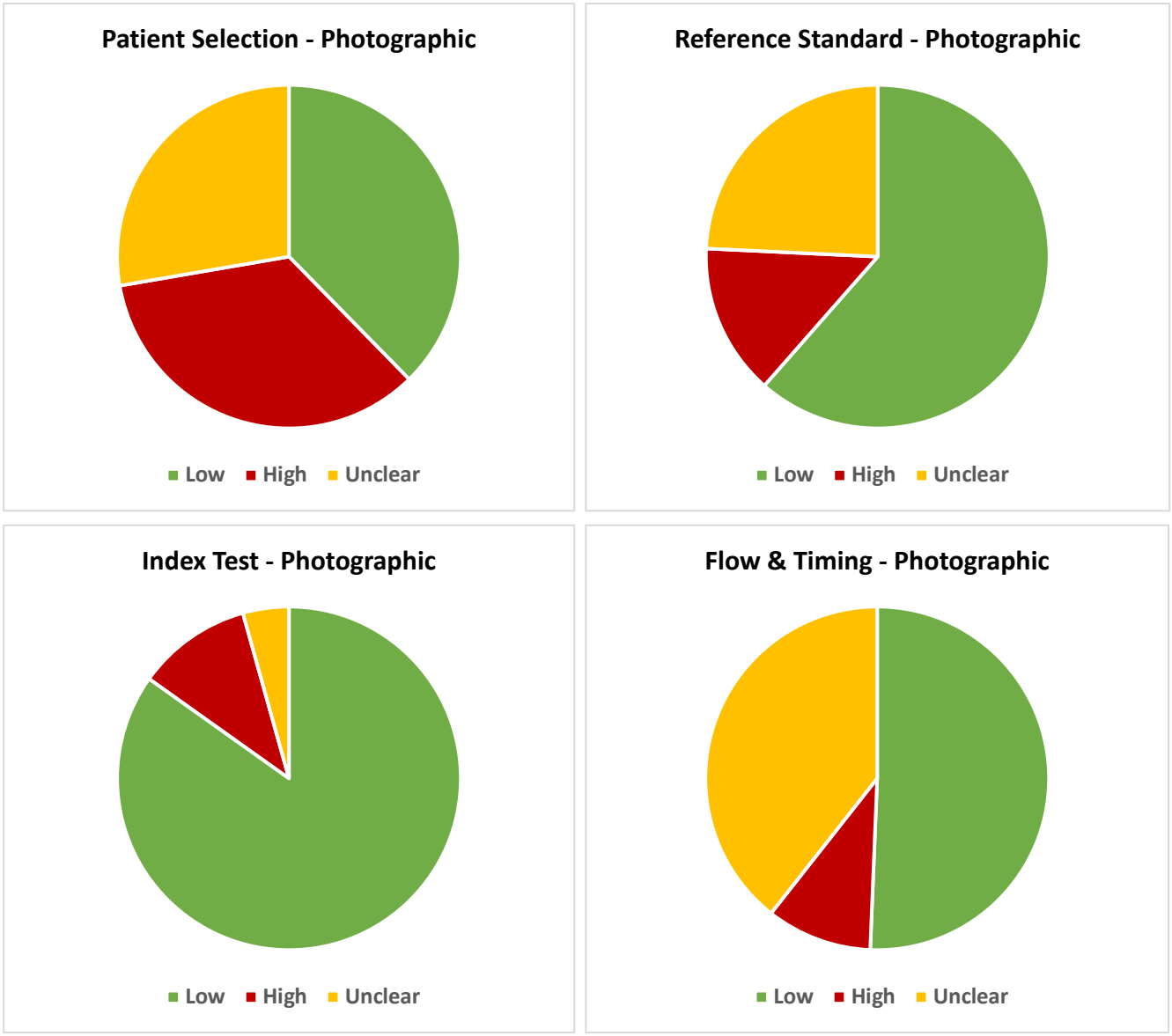


Figure 4: Pie charts demonstrating the risk of bias amongst photographic images studies, as assessed through QUADAS

Table 8: Quality Assessment and Adherence to QUADAS in Systematic Reviews of Diagnostic Accuracy of Artificial Intelligence in Pathology

Study	Modality	Quality assessment?	QUADAS?	Modifications	Other Tools Used	Table
Azam 2020	Histology samples	Yes	Yes	No	No	Yes
Mahmood 2020	Histology samples	Yes	Yes	Yes – modified QUADAS	No	No
Azer 2019	Histology samples	No	No	No	No	No

Perceived Limitations

13 studies reported an inability to provide systematic quality assessment or evaluate certain biases as a limitation in their study (Supplementary Table 2). Specifically, these included concerns around size and quality of the dataset, including its real-world clinical applicability; for example including a whole tissue section instead of a portion of interest only[23] and providing samples from multiple centres across different demographic populations to improve generalisability of the model. Appropriate separation of data set into training, validation and test sets without overlap was also highlighted as an area needing evaluation, as overlap between datasets would lead to higher accuracy rates. Eight reviews modified or tailored pre-existing quality assessment tools to customise it to the methodologies and types of studies as reported above.

Discussion

This study demonstrates that formal quality appraisal and risk of bias assessment is not uniformly applied in AI-based diagnostic accuracy systematic reviews. Despite being considered a prerequisite, only 75% of studies performed any form of quality assessment; with 56% of reviews opting for the QUADAS tool. Despite the most commonly used tool in this field, the use of both new and modified tools (e.g. RQS tool) suggests that the current instruments are poorly suited for AI centred diagnostic accuracy studies.

In the patient selection domain, 113 studies (26.7% of studies) were deemed high risk and an additional 30.7% of studies were deemed to be of unclear risk of bias (n=130). This risk was greatest in studies reporting on photographic images, where 35% of studies were at high risk of bias (Figure 5). Factors leading to high risk of bias in patient selection include poor patient sampling technique as well as inappropriate exclusion of data on a patient or feature level. As AI algorithms rely on previously seen data to identify patterns and generate results, the accuracy of the algorithm will be directly related to the accuracy of the input data. Consequently, artefacts, inaccuracies or biases in input data can be perpetuated and augmented by the model and under-representation of certain factors or demographics may result in inferior algorithm performance.[59] Therefore, inappropriate representation of patient demographics or socioeconomic factors may also manifest in the algorithm output as discriminate results. This type of bias may be aggravated in photographic images where utilising images or data derived from a specific demographic may create blind spots in the AI algorithm amplifying racial biases.[60] For example, employing an AI model to detect dermatological abnormalities on dark skin resulted in higher rates of missed diagnoses further increasing the disparity in diagnosis.[61,62]

In addition to a lack of diversity within the input data, there are several other sources of AI-specific biases including historical bias, representation bias, evaluation bias, aggregation bias, population bias and sampling bias which are discussed in detail by Mehrabi et al and Simpson's Paradox, detailed further below.[63] Evaluation bias pertains to the use of inappropriate benchmarks for evaluation of the algorithm. Representation bias arises from misdefining and consequently mis-sampling the population. For example, amongst algorithms trained on United States data, the majority overwhelmingly featured data from New York, Massachusetts and California and included little to no data on the remainder of the country.[64] These states do not represent the United States as a whole, and have considerable differences in socioeconomic, educational, racial and cultural characteristics, resulting in a lack of generalisability of the algorithm to the remainder of the country. Historical biases refer to pre-existing biases in the input data. Of note, historical biases may be further perpetuated and amplified by AI algorithms. Additionally, using a proxy that appears to correlate with the outcome of interest may be a significant source of bias. For example, a study of risk-based payment algorithms utilising healthcare costs as a proxy for health needs was found to falsely conclude Black patients were healthier than similarly comorbid White patients, despite the

algorithm specifically excluding data on race. Owing to inequal access to care, health-related spending was similar for a White patient and a Black patient with more severe illness, and utilising cost as proxy therefore created racial biases.[65] Other concerns include the methods around use of AI outputs in clinical practice and improper implementation can lead to implicit biases.

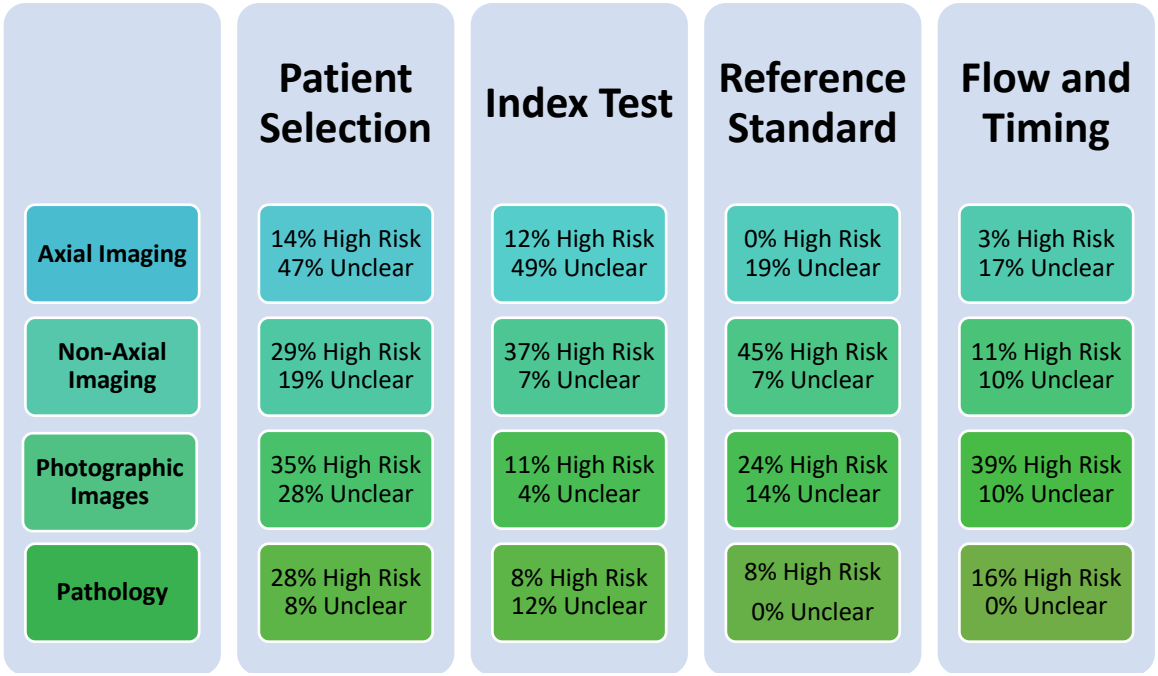
Additionally, AI-based diagnostics require rigorous datasets representative of real-world characteristics to produce reliable and generalisable diagnostic results. Therefore, inappropriate exclusion of participants and/or features can affect the interpretation of AI results in a more significant way compared to conventional analysis and contribute to bias in patient selection. Excluding pathologies that may be similar in characteristics or have overlapping features to the diagnoses of interest may lead to overestimation of the algorithm's diagnostic accuracy as well as low generalisability and clinical utility of the algorithm. For example, exclusion of patients with inflammatory bowel disease when determining diagnostic accuracy of colonoscopic detection of polyps reduces the ability of the algorithm to discriminate between benign polyps and more serious pathologies.[66] Another example includes medical photographs where images may be more likely to be excluded due to poor quality. Excluding blurry or out-of-focus images may lead to falsely elevated diagnostic accuracy rates and also does not reflect real-world situations thereby reducing study applicability and clinical value. Finally, in comparison to conventional index tests which require description of sampling methods on a patient only, AI models also require the description of sampling input level data[67]; insufficient description of this may have led to considerable studies presenting an unclear risk of bias.

Within the index test domain, both axial and non-axial imaging studies had a high risk of bias. The index test domain pertains to the development and validation of the AI algorithm and interpretation of the generated output. First, distributional shifts between the training, validation and testing datasets can result in the algorithm producing incorrect results with confidence. These shifts can also lead to inaccurate conclusions about the precision of the algorithm if the algorithm is tested inappropriately on a patient cohort for which it was not trained.[68] Secondly, overlapping datasets can overestimate diagnostic accuracy in comparison to using external validation data. Thirdly, given the heterogenous nature of large datasets utilised for AI, there is increased possibility of confounding factors amongst the data. If the model does not appropriately address these causal relations between different factors (i.e., subgroups within the dataset as a result of these confounding factors), this can lead to Simpson's Paradox, which arises from aggregated analysis of heterogenous data comprised of multiple subgroups - separating the dataset into different groups based on the confounding variables provides a different result compared to analysing all the data together.[63] Finally, the size of the dataset is particularly important for AI models as small datasets are more likely to provide lower diagnostic accuracy and also result in poor generalisability of the results.[69] Additionally, a lack of exposure to multiple manifestations of a pathology can result in the

'Frame Problem' whereby seemingly obvious diagnoses may be missed by the model simply due to a inexperience. Specific signalling questions addressing these potential areas of concern may be useful in identifying and characterising sources of bias and determining generalisability of the AI model.

48 studies (11.4%) posed a high risk of bias in the reference standard domain. Though non-axial imaging studies appeared to be disproportionately at higher risk of bias in this domain, all studies resulted from one systematic review.[42] Though overall low risk, this domain contains several potential sources of bias for AI-specific studies of diagnostic accuracy. Determination of an appropriate reference standard, or 'ground truth' for training of models, requires consideration of the best available evidence and may involve amalgamating clinical, radiological and laboratory data.[69] Comparison of AI against a human reference standard can be utilised if it is considered the gold standard though should be avoided as a sole reference standard if an alternative test providing higher sensitivity and specificity is feasible. For example, 32 of 33 studies in the systematic review by Harris et al. were deemed to be at high risk of bias primarily due to the reference standard comprising human interpretation of the chest x-ray without use of sputum culture confirmation.[42] Where comparison is against a human reference standard, the number of operators, their experience and presence of interobserver variability should be clearly detailed. Ideally, the reference standard should include multiple annotations from different experts to reduce subjectivity and account for interobserver variability.[23] This is particularly important in the context of AI given its potential capabilities in detecting disease more accurately than human operators.[1] Furthermore, AI may also be capable of detecting subtle changes indicative of a diagnosis through recognition of patterns not detectable by human operators, for example deriving cardiovascular risk from retinal images, identification of individuals with atrial fibrillation from their ECGs taken during sinus rhythm and identifying stromal features associated with breast cancer survival.[70–72] Therefore, a combination of investigations including repeat tests undertaken at different time points may be required as a reference standard, particularly as a greater number of prospective studies emerge.

Figure 5: Summary of Risk of Bias Across the QUADAS Domains



Finally, the flow and timing domain considers interval between the index test and reference standard, similarities in reference standard evaluation amongst all patients and inclusion of all patients in the final analysis. Within this domain, studies performed reasonably well with only 37 studies (8.8%) recorded as high risk of bias. However, methodologies relating to study flow and standards of timing vary in AI-based studies representing a different risk of bias. For example, neuropsychiatric studies utilising AI have been able to detect the presence of early cognitive changes or aid diagnosis of psychiatric disorders through identification of otherwise indiscernible changes in structural or functional neuroimaging.[24,40,54] In mild or initial stages of disease, AI may actually be more discriminant than the reference standard in identifying early variations or subtle patterns.[24,41] Therefore, the timing of the reference standard in relation to the index test is imperative and may need to be scheduled at a later date to ensure the diagnosis reflected by the reference standard is accurate. Furthermore, use of different reference standard between positive and negative cases may pose further challenges in AI-based studies. For example, histology confirmation is often utilised as part of the reference standard for confirming malignancy but obtaining biopsies from clearly benign lesions poses ethical and practical challenges, thereby necessitating the use of alternative confirmatory tests.[48] However, utilising vastly different reference standards such as follow-up alone in comparison to histology may result in verification bias i.e. false negatives may actually be classed as true negatives and inflate estimates of accuracy. In these cases where an alternative reference standard is required, utilising an investigation with high negative predictive value such as clinical follow-up with a PET scan to rule out malignancy may be suitable.[73] However in AI-based studies, additional considerations have to be given for similarities between the ground truth used to train the model and the reference standard

used to validate and test the model. If there are considerable disparities between the two, the model may be erroneously be deemed inadequate.

Perceived limitations of current quality assessment tools highlight the need for an AI-specific guideline to evaluate diagnostic accuracy studies. Algorithm and input data quality, real-world clinical applicability and algorithm generalisability are important sources of bias that need to be addressed in an adapted AI-specific tool. Quality assessment tools similar to QUADAS are currently being modified to match the evolving landscape of research. For example, STARD (Standards for Reporting of Diagnostic Accuracy Studies), is currently being extended to develop the STARD-AI guidelines to specifically appraise AI based diagnostic accuracy studies.[67] Additionally, AI extensions to TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) and CONSORT (Consolidated Standards of Reporting Trials) have been published, and SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials) is in progress.[74–76]

Conclusions

This review demonstrates incomplete uptake of quality assessment tools in AI centred diagnostic accuracy reviews and highlights variations in AI-specific methodological aspects and reporting across all domains of QUADAS in particular. These factors include generalisability and diversity in patient selection, development of training, validation and testing datasets, as well as definition and evaluation of the reference standard and comparison with human performance. When evaluating study quality, potential biases and applicability of AI diagnostic accuracy studies, it is imperative that systematic reviews consider these factors. Whilst the QUADAS-2 tool explicitly recognises the difficulty in developing a tool generalisable to all studies across all specialties and topics and proposes the author modifies the signalling questions as needed, it is essential to further define these questions for AI studies given complexities in methodology.

Inadequate reporting may create barriers to clinical implementation of AI-based diagnostic tools and also result in a lack of reproducibility, thereby leading to an inability to validate models in other geographical or demographical areas and hindering wider use in clinical practice. Inherent methodological differences in AI-based tools together with the consequences of inadequate reporting of studies and low adherence to QUADAS in these systematic reviews highlights the need for an AI specific framework. We propose the creation of a QUADAS-AI extension emulating the successful development of AI extensions to other quality assessment tools.[67,74,75] QUADAS-AI and STARD-AI may be employed in parallel to harmonise evaluation of diagnostic accuracy studies. The adoption of a robust and accepted instrument to assess the quality of primary diagnostic accuracy AI studies for integration within a systematic review can offer an evidence-base to safely translate AI tools into a real world setting that can maximise the benefits of AI for the future of medical diagnostics and care.

References

1. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
2. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6
3. Yamada M, Saito Y, Imaoka H, et al. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci Rep*. 2019;9(1):1-9. doi:10.1038/s41598-019-50567-5
4. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digit Med*. 2019;2(1):1-10. doi:10.1038/s41746-019-0112-2
5. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:1-13. doi:10.1186/1471-2288-3-25
6. Whiting PF, Rutjes AWS, Westwood ME, et al. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009
7. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141
8. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*. 2009;339(7716):332-336. doi:10.1136/bmj.b2535
9. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358. doi:10.1136/bmj.j4008
10. Ursprung S, Beer L, Bruining A, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma—a systematic review and meta-analysis. *Eur Radiol*. 2020;30(6):3558-3566. doi:10.1007/s00330-020-06666-3
11. Cho SJ, Sunwoo L, Baik SH, Bae YJ, Choi BS, Kim JH. Brain metastasis detection using machine learning: a systematic review and meta-analysis. *Neuro Oncol*. October 2020:1-12. doi:10.1093/neuonc/noaa232
12. Pellegrini E, Ballerini L, Hernandez M del CV, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's Dement Diagnosis, Assess Dis Monit*. 2018;10:519-535. doi:10.1016/j.dadm.2018.07.004
13. Langerhuizen DWG, Janssen SJ, Mallee WH, et al. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. *Clin Orthop Relat Res*. 2019;477(11):2482-2491. doi:10.1097/CORR.0000000000000848
14. Groot OQ, Bongers MER, Ogink PT, et al. Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review. *Clin Orthop Relat Res*. 2020;478(12):2751-2764.

doi:10.1097/CORR.0000000000001360
15. Nayantara PV, Kamath S, Manjunath KN, Rajagopal K V. Computer-aided diagnosis of liver lesions using CT images: A systematic review. *Comput Biol Med*. 2020;127. doi:10.1016/j.compbiomed.2020.104035
16. Crombé A, Fadli D, Italiano A, Saut O, Buy X, Kind M. Systematic review of sarcomas radiomics studies: Bridging the gap between concepts and clinical applications? *Eur J Radiol*. 2020;132. doi:10.1016/j.ejrad.2020.109283
17. Kunze KN, Rossi DM, White GM, et al. Diagnostic Performance of Artificial Intelligence for Detection of Anterior Cruciate Ligament and Meniscus Tears: A Systematic Review. *Arthrosc - J Arthrosc Relat Surg*. September 2020. doi:10.1016/j.arthro.2020.09.012
18. Bang CS, Lee JJ, Baik GH. Artificial intelligence for the prediction of helicobacter pylori infection in endoscopic images: Systematic review and meta-analysis of diagnostic test accuracy. *J Med Internet Res*. 2020;22(9). doi:10.2196/21983
19. Azam AS, Miligy IM, Kimani PKU, et al. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *J Clin Pathol*. 2020;0:1-8. doi:10.1136/jclinpath-2020-206764
20. Xu L, Gao J, Wang Q, et al. Computer-Aided Diagnosis Systems in Diagnosing Malignant Thyroid Nodules on Ultrasonography: A Systematic Review and Meta-Analysis. *Eur Thyroid J*. 2020;9(4):186-193. doi:10.1159/000504390
21. Li Y, Zhang Z, Dai C, Dong Q, Badrigilan S. Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: A systematic review and meta-analysis. *Comput Biol Med*. 2020;123. doi:10.1016/j.compbiomed.2020.103898
22. Mohan BP, Khan SR, Kassab LL, et al. High pooled performance of convolutional neural networks in computer-aided diagnosis of GI ulcers and/or hemorrhage on wireless capsule endoscopy images: a systematic review and meta-analysis. *Gastrointest Endosc*. 2020;93(2):356-364.e4. doi:10.1016/j.gie.2020.07.038
23. Mahmood H, Shaban M, Indave BI, Santos-Silva AR, Rajpoot N, Khurram SA. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. *Oral Oncol*. 2020;110. doi:10.1016/j.oraloncology.2020.104885
24. Steardo L, Carbone EA, de Filippis R, et al. Application of Support Vector Machine on fMRI Data as Biomarkers in Schizophrenia Diagnosis: A Systematic Review. *Front Psychiatry*. 2020;11:588. doi:10.3389/fpsyt.2020.00588
25. Hassan C, Spadaccini M, Iannone A, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointest Endosc*. 2021;93(1):77-85.e6. doi:10.1016/j.gie.2020.06.059
26. Yang S, Yin B, Cao W, Feng C, Fan G, He S. Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. *Clin Radiol*. 2020;75(9):713.e17-713.e28. doi:10.1016/j.crad.2020.05.021
27. Lui TKL, Tsui VWM, Leung WK. Accuracy of artificial intelligence–assisted detection of upper GI lesions: a systematic review and meta-analysis. *Gastrointest Endosc*. 2020;92(4):821-830.e9. doi:10.1016/j.gie.2020.06.034
28. Wang S, Zhang Y, Lei S, et al. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: A systematic review and meta-analysis of diagnostic test accuracy. *Eur J Endocrinol*. 2020;183(1):41-49. doi:10.1530/EJE-19-0968

29. Ninatti G, Kirienko M, Neri E, Sollini M, Chiti A. Imaging-based prediction of molecular therapy targets in NSCLC by radiogenomics and AI approaches: A systematic review. *Diagnostics*. 2020;10(6). doi:10.3390/diagnostics10060359
30. Li J, Sang T, Yu W-H, et al. The value of S-Detect for the differential diagnosis of breast masses on ultrasound: a systematic review and pooled meta-analysis. *Med Ultrason*. 2020;22(2):211. doi:10.11152/mu-2402
31. Soffer S, Klang E, Shimon O, et al. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointest Endosc*. 2020;92(4):831-839.e8. doi:10.1016/j.gie.2020.04.039
32. Islam MM, Yang HC, Poly TN, Jian WS, (Jack) Li YC. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Comput Methods Programs Biomed*. 2020;191:105320. doi:10.1016/j.cmpb.2020.105320
33. Iannattone PA, Zhao X, VanHouten J, Garg A, Huynh T. Artificial Intelligence for Diagnosis of Acute Coronary Syndromes: A Meta-analysis of Machine Learning Approaches. *Can J Cardiol*. 2020;36(4):577-583. doi:10.1016/j.cjca.2019.09.013
34. Lui TKL, Guo CG, Leung WK. Accuracy of artificial intelligence on histology prediction and detection of colorectal polyps: a systematic review and meta-analysis. *Gastrointest Endosc*. 2020;92(1):11-22.e6. doi:10.1016/j.gie.2020.02.033
35. Islam MM, Poly TN, Walther BA, Yang HC, Li Y-C (Jack). Artificial Intelligence in Ophthalmology: A Meta-Analysis of Deep Learning Models for Retinal Vessels Segmentation. *J Clin Med*. 2020;9(4):1018. doi:10.3390/jcm9041018
36. Halder A, Dey D, Sadhu AK. Lung Nodule Detection from Feature Engineering to Deep Learning in Thoracic CT Images: a Comprehensive Review. *J Digit Imaging*. 2020;33(3):655-677. doi:10.1007/s10278-020-00320-6
37. Murtagh P, Greene G, O'Brien C. Current applications of machine learning in the screening and diagnosis of glaucoma: A systematic review and Meta-analysis. *Int J Ophthalmol*. 2020;13(1):149-162. doi:10.18240/ijo.2020.01.22
38. Azer SA. Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: A systematic review. *World J Gastrointest Oncol*. 2019;11(12):1218-1230. doi:10.4251/wjgo.v11.i12.1218
39. Li D, Vilmun BM, Carlsen JF, et al. The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from LIDC-IDRI: A systematic review. *Diagnostics*. 2019;9(4). doi:10.3390/diagnostics9040207
40. Moon SJ, Hwang J, Kana R, Torous J, Kim JW. Accuracy of Machine Learning Algorithms for the Diagnosis of Autism Spectrum Disorder: Systematic Review and Meta-Analysis of Brain Magnetic Resonance Imaging Studies. *JMIR Ment Heal*. 2019;6(12):e14108. doi:10.2196/14108
41. Jo T, Nho K, Saykin AJ. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front Aging Neurosci*. 2019;11:220. doi:10.3389/fnagi.2019.00220
42. Harris M, Qi A, Jeagal L, et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. *PLoS One*. 2019;14(9). doi:10.1371/journal.pone.0221339
43. Sarmento RM, Vasconcelos FFX, Filho PPR, Wu W, De Albuquerque VHC. Automatic Neuroimage Processing and Analysis in Stroke - A Systematic Review. *IEEE Rev*

Biomed Eng. 2020;13:130-155. doi:10.1109/RBME.2019.2934500
44. Zhao WJ, Fu LR, Huang ZM, Zhu JQ, Ma BY, Tarantino G. Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid nodules on ultrasound: A systematic review and meta-analysis. *Med (United States)*. 2019;98(32). doi:10.1097/MD.0000000000016379
45. De Filippis R, Carbone EA, Gaetano R, et al. Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: A systematic review. *Neuropsychiatr Dis Treat*. 2019;15:1605-1627. doi:10.2147/NDT.S202418
46. Nielsen KB, Lautrup ML, Andersen JKH, Savarimuthu TR, Grauslund J. Deep Learning–Based Algorithms in Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance. *Ophthalmol Retin*. 2019;3(4):294-304. doi:10.1016/j.oret.2018.10.014
47. Pehrson LM, Nielsen MB, Lauridsen CA. Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the LIDC-IDRI database: A systematic review. *Diagnostics*. 2019;9(1). doi:10.3390/diagnostics9010029
48. Marka A, Carter JB, Toto E, Hassanpour S. Automated detection of nonmelanoma skin cancer using digital images: A systematic review. *BMC Med Imaging*. 2019;19(1):21. doi:10.1186/s12880-019-0307-7
49. Ferrante di Ruffano L, Takwoingi Y, Dinnes J, et al. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. *Cochrane Database Syst Rev*. 2018;2018(12). doi:10.1002/14651858.CD013186
50. Chuchu N, Takwoingi Y, Dinnes J, et al. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. *Cochrane Database Syst Rev*. 2018;2018(12). doi:10.1002/14651858.CD013192
51. Nguyen A V., Blears EE, Ross E, Lall RR, Ortega-Barnett J. Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: A systematic review and meta-analysis. *Neurosurg Focus*. 2018;45(5):E5. doi:10.3171/2018.8.FOCUS18325
52. McCarthy J, Collins DL, Ducharme S. Morphometric MRI as a diagnostic biomarker of frontotemporal dementia: A systematic review to determine clinical applicability. *NeuroImage Clin*. 2018;20:685-696. doi:10.1016/j.nicl.2018.08.028
53. Farzandipour M, Nabovati E, Saeedi S, Fakharian E. Fuzzy decision support systems to diagnose musculoskeletal disorders: A systematic literature review. *Comput Methods Programs Biomed*. 2018;163:101-109. doi:10.1016/j.cmpb.2018.06.002
54. Bruin W, Denys D, van Wingen G. Diagnostic neuroimaging markers of obsessive-compulsive disorder: Initial evidence from structural and functional MRI studies. *Prog Neuro-Psychopharmacology Biol Psychiatry*. 2019;91:49-59. doi:10.1016/j.pnpbp.2018.08.005
55. Senders JT, Arnaout O, Karhade A V., et al. Natural and artificial intelligence in neurosurgery: A systematic review. *Clin Neurosurg*. 2018;83(2):181-192. doi:10.1093/neuros/nyx384
56. Smith A, López-Solà M, McMahon K, Pedler A, Sterling M. Multivariate pattern analysis utilizing structural or functional MRI—In individuals with musculoskeletal pain and healthy controls: A systematic review. *Semin Arthritis Rheum*. 2017;47(3):418-431. doi:10.1016/j.semarthrit.2017.06.005
57. Sprockel J, Tejeda M, Yate J, Diaztagle J, González E. Intelligent systems tools in the

diagnosis of acute coronary syndromes: A systemic review. *Arch Cardiol Mex*. 2018;88(3):178-189. doi:10.1016/j.acmx.2017.03.002
58. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. *Br J Dermatol*. 2009;161(3):591-604. doi:10.1111/j.1365-2133.2009.09093.x
59. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*. 2020;117(23):12592-12594. doi:10.1073/pnas.1919012117
60. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
61. Kamulegeya LH, Okello M, Bwanika JM, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. *bioRxiv*. October 2019:826057. doi:10.1101/826057
62. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatology*. 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348
63. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. *A Survey on Bias and Fairness in Machine Learning*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed March 14, 2021.
64. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA - J Am Med Assoc*. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067
65. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-)*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
66. Gross S, Trautwein C, Behrens A, et al. Computer-based classification of small colorectal polyps by using narrow-band imaging with optical magnification. *Gastrointest Endosc*. 2011;74(6):1354-1359. doi:10.1016/j.gie.2011.08.001
67. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med*. 2020;26(6):807-808. doi:10.1038/s41591-020-0941-1
68. Challen R, Denny J, Pitt M, Gompels L. Challen R, et al. BMJ Qual Artificial intelligence, bias and clinical safety. *Saf*. 2019;28:231-237. doi:10.1136/bmjqs-2018-008370
69. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295(1):4-15. doi:10.1148/radiol.2020192224
70. Beck AH, Sangoi AR, Leung S, et al. Imaging: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*. 2011;3(108). doi:10.1126/scitranslmed.3002564
71. Poplin R, Varadarajan A V., Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0
72. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet*. 2019;394(10201):861-867. doi:10.1016/S0140-6736(19)31721-0
73. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference

standard. *J Clin Epidemiol*. 62:797-806. doi:10.1016/j.jclinepi.2009.02.005
74. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
75. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6
76. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7

Figures

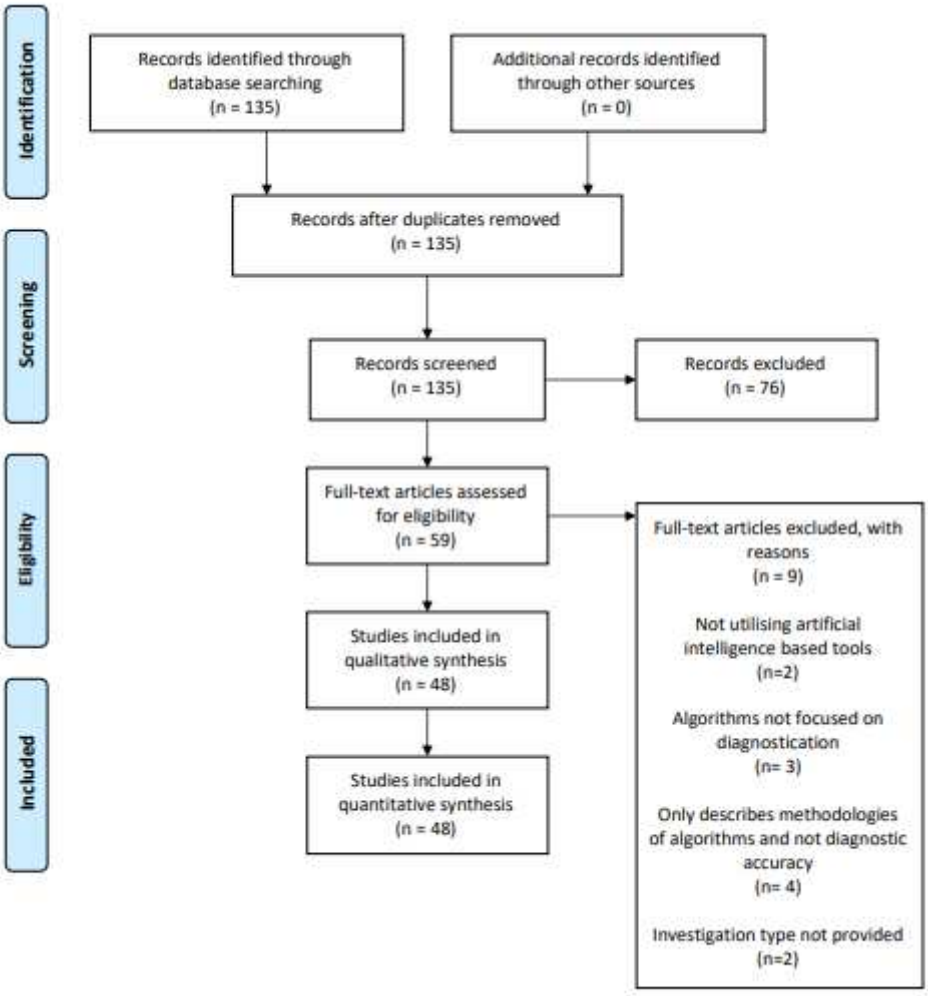


Figure 1

PRISMA Guidelines

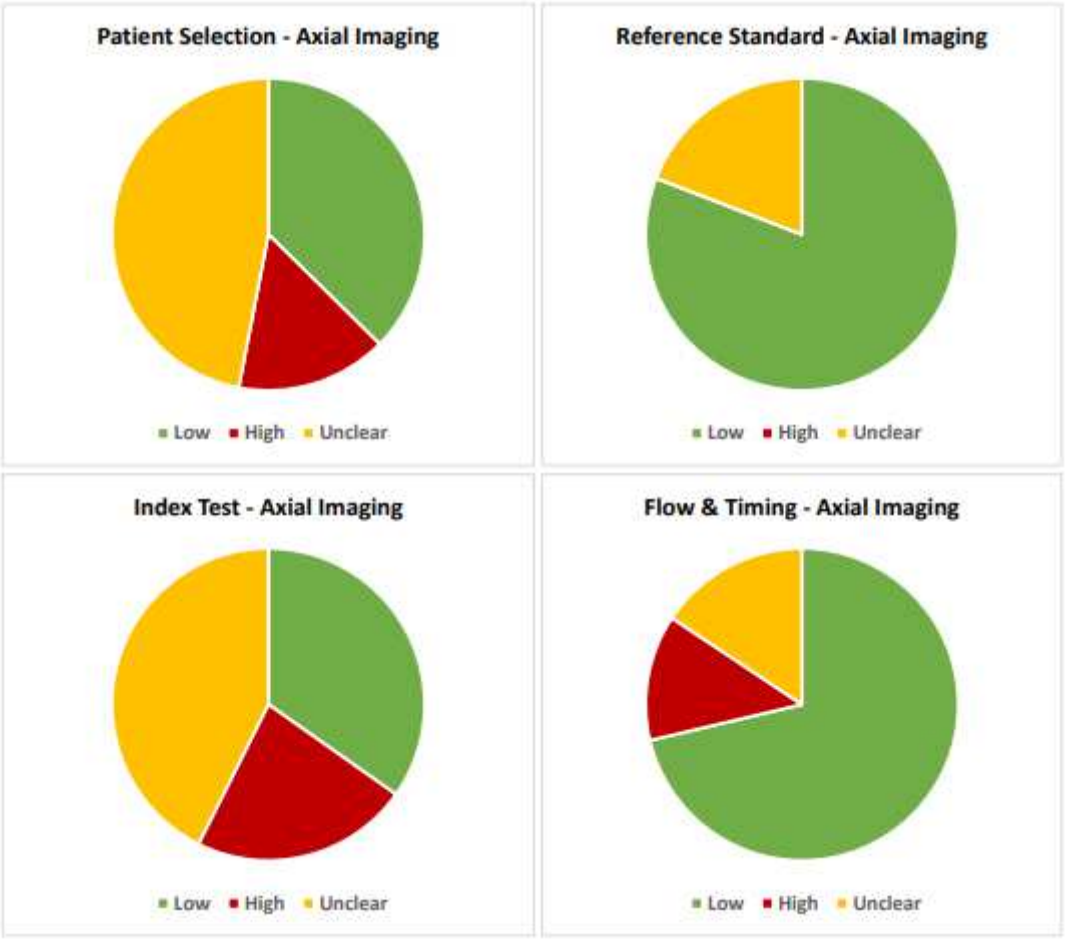


Figure 2

Pie charts demonstrating the risk of bias amongst axial imaging studies, as assessed through QUADAS

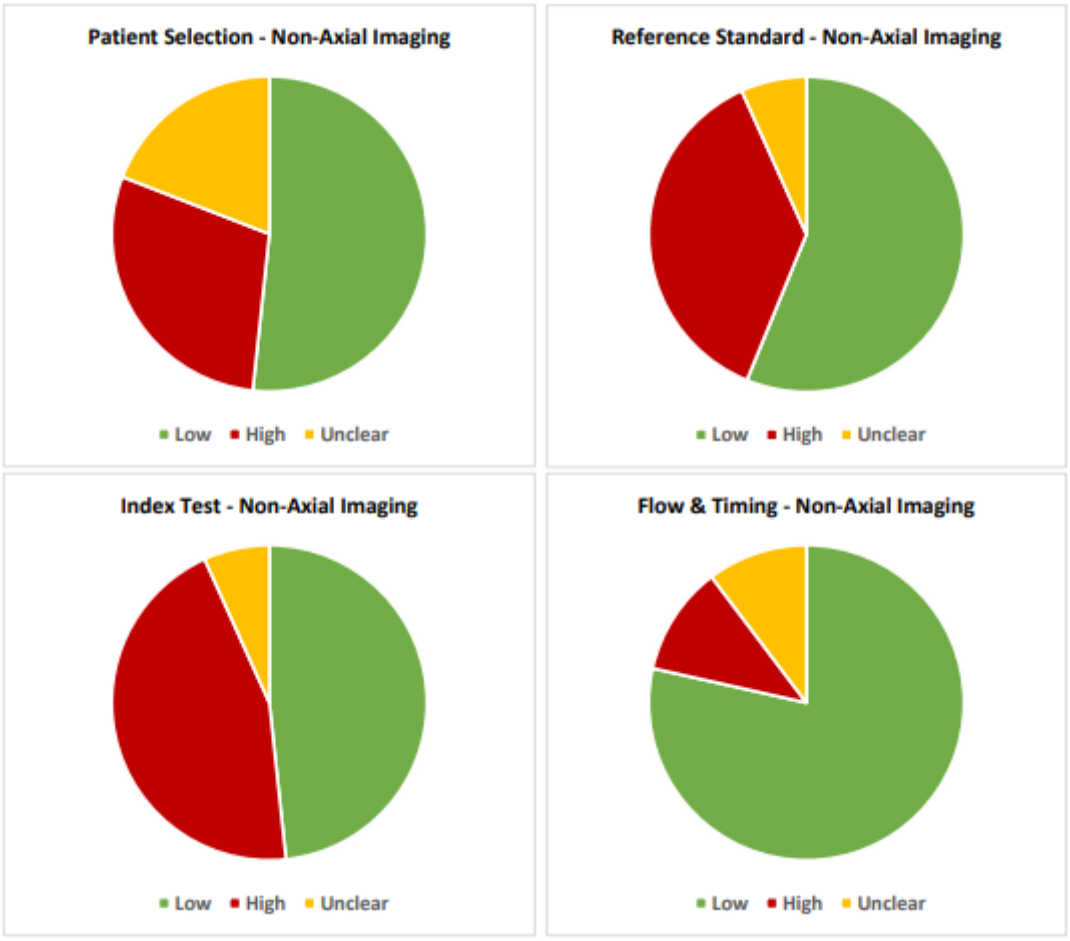


Figure 3

Pie charts demonstrating the risk of bias amongst non-axial imaging studies, as assessed through QUADAS

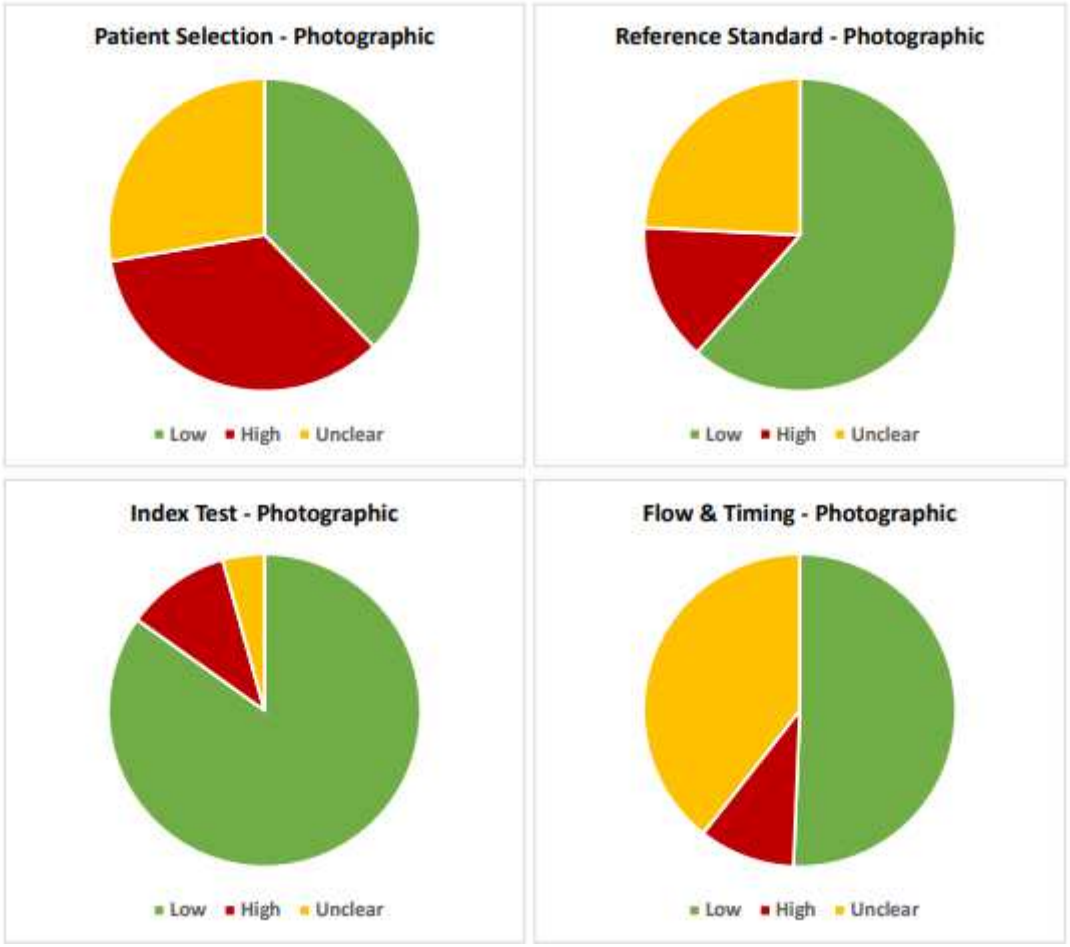


Figure 4

Pie charts demonstrating the risk of bias amongst photographic images studies, as assessed through QUADAS

	Patient Selection	Index Test	Reference Standard	Flow and Timing
Axial Imaging	14% High Risk 47% Unclear	12% High Risk 49% Unclear	0% High Risk 19% Unclear	3% High Risk 17% Unclear
Non-Axial Imaging	29% High Risk 19% Unclear	37% High Risk 7% Unclear	45% High Risk 7% Unclear	11% High Risk 10% Unclear
Photographic Images	35% High Risk 28% Unclear	11% High Risk 4% Unclear	24% High Risk 14% Unclear	39% High Risk 10% Unclear
Pathology	28% High Risk 8% Unclear	8% High Risk 12% Unclear	8% High Risk 0% Unclear	16% High Risk 0% Unclear

Figure 5

Summary of Risk of Bias Across the QUADAS Domains

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFiles.docx