

1

2

3 Diagnostic test accuracy of artificial intelligence in screening for referable diabetic  
4 retinopathy in real-world settings: A systematic review and meta-analysis

5

6

7 Holijah Uy<sup>1\*</sup>, Christopher Fielding<sup>2</sup>, Ameer Hohlfeld<sup>3</sup>, Eleanor Ochodo<sup>4,5</sup>,  
8 Abraham Opare<sup>1</sup>, Elton Mukonda<sup>2</sup>, Deon Minnies<sup>1¶</sup>, Mark E Engel<sup>3,6\*¶</sup>

9

10

11

12 <sup>1</sup> Community Eye Health Institute, Faculty of Health Sciences, University of Cape Town,  
13 South Africa

14 <sup>2</sup> Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health  
15 Sciences, University of Cape Town, South Africa

16 <sup>3</sup> South African Medical Research Council, South Africa

17 <sup>4</sup> Centre for Global Health Research, Kenya Medical Research Institute, Kenya

18 <sup>5</sup> Department of Global Health, Stellenbosch University, South Africa

19 <sup>6</sup> Department of Medicine, University of Cape Town, South Africa

20

21 \* Corresponding authors

22 E-mail: [mark.engel@uct.ac.za](mailto:mark.engel@uct.ac.za) (MEE); [uyxhol001@myuct.ac.za](mailto:uyxhol001@myuct.ac.za) (HU)

23

24 ¶DM and MEE are Joint Senior Authors

## 25 **Abstract**

26 Studies on artificial intelligence (AI) in screening for diabetic retinopathy (DR) have shown promising  
27 results in addressing the mismatch between the capacity to implement DR screening and the increasing  
28 DR incidence; however, most of these studies were done retrospectively. This review sought to evaluate  
29 the diagnostic test accuracy (DTA) of AI in screening for referable diabetic retinopathy (RDR) in real-  
30 world settings. We searched CENTRAL, PubMed, CINAHL, Scopus, and Web of Science on 9  
31 February 2023. We included prospective DTA studies assessing AI against trained human graders  
32 (HGs) in screening for RDR in patients living with diabetes. synthesis Two reviewers independently  
33 extracted data and assessed methodological quality against QUADAS-2 criteria. We used the  
34 hierarchical summary receiver operating characteristics (HSROC) model to pool estimates of sensitivity  
35 and specificity and, forest plots and SROC plots to visually examine heterogeneity in accuracy  
36 estimates. Finally, we conducted sensitivity analyses to explore the effects of studies deemed to possibly  
37 affect the quality of the studies. We included 15 studies (17 datasets: 10 patient-level analysis  
38 (N=45,785), and 7 eye-level analysis (N=15,390). Meta-analyses revealed a pooled sensitivity of  
39 95.33%(95% CI: 90.60-100%) and specificity of 92.01%(95% CI: 87.61-96.42%) for patient-level  
40 analysis; for the eye-level analysis, pooled sensitivity was 91.24% (95% CI: 79.15-100%) and  
41 specificity, 93.90% (95% CI: 90.63-97.16%). Subgroup analyses did not provide variations in the  
42 diagnostic accuracy of country classification and DR classification criteria; however, a moderate  
43 increase was observed in diagnostic accuracy at the primary-level and, a minimal decrease in the  
44 tertiary-level healthcare settings. Sensitivity analyses did not show any variations in studies that  
45 included diabetic macular edema in the RDR definition, nor in studies with  $\geq 3$  HGs. This review  
46 provides evidence, for the first time from prospective studies, for the effectiveness of AI in screening  
47 for RDR, in real-world settings.

48

49

## 50 **Introduction**

51 Diabetic retinopathy (DR) is the most common and specific complication of diabetes mellitus in the  
52 working age group [1]. In 2020, the number of adults with DR was estimated to be 103.12 million,  
53 which is expected to be 129.84 million by 2030 and 160.50 million by 2045 [1]. Along with an  
54 increasing incidence of DR, the number of people with vision impairment and blindness also increases.  
55 Without early intervention, the incidence of blindness due to DR will continue to rise as the number of  
56 people getting diabetes increases. Thus, DR has become a global public health concern, compelling  
57 researchers and health practitioners to continuously develop strategies to prevent and treat DR.

58 Diabetic retinopathy can be asymptomatic for years, even at advanced stages [2]. Thus, early-stage  
59 detection of DR is crucial to provide timely treatment and management. For that reason, DR screening  
60 programmes are being implemented in public health settings through population-based or opportunistic  
61 screening. Diabetic retinopathy screening aims to distinguish between patients who need a referral,  
62 termed referable DR (RDR), for ophthalmological intervention from those who can continue annual  
63 routine eye care services [3]. Referable DR can be classified as moderate nonproliferative DR (NPDR)  
64 or worse and/or diabetic macular edema (DME). Those with RDR must be referred within three months  
65 to one year, depending on the resource settings [4].

66 Currently, local and international programmes combatting DR are facing a significant crisis due to the  
67 increasing prevalence of diabetes. This influx has outpaced the development of healthcare services and  
68 screening programmes for preventing DR [5]. According to a systematic review by Piyasena et al. [6],  
69 aside from the high cost of services and lack of infrastructure for retinal imaging and training  
70 programmes, one of the major barriers to DR screening is the lack of skilled human resources, especially  
71 in the lower- and middle-income countries.

72 Artificial intelligence has shown to be a promising solution to these challenges by functioning in an  
73 autonomous mode. Through deep learning algorithms, AI can be used to detect the presence and severity  
74 of DR in real-time. However, it is crucial that these tools should have high diagnostic accuracy and  
75 good performance before being implemented in various healthcare settings. The UK National Institute  
76 for Clinical Excellence (NICE) Guidelines stated that DR screening programmes should use screening

77 tools with a sensitivity of  $\geq 80\%$ , specificity of  $\geq 95\%$ , and a technical failure rate of  $\leq 5\%$  [7].  
78 Meanwhile, the St Vincent Declaration of 2005 suggested that systematic DR screening programmes  
79 should aim for a sensitivity of  $\geq 80\%$  and a specificity of  $\geq 90\%$  with an acceptable coverage of  $\geq 80\%$   
80 [8].  
81 In recent years, retrospective validation studies have shown AI to have high diagnostic accuracy in  
82 detecting DR; that is, AI is equally good or even better than human graders (HGs). Studies done in real-  
83 world settings using prospective data collection have also demonstrated robust performance [9];  
84 however, these studies are fewer than those done in a retrospective manner, and the true utility of AI  
85 systems in DR screening will only be better understood through prospective studies, as performance is  
86 likely to be affected when dealing with real-world data that is different from the data used for algorithm  
87 training [10, 11]. Moreover, prospective studies, with pre-established protocols, allow them to be more  
88 robust and generalisable, and exhibit the true impact on system usability in real-world settings.  
89 Therefore, we conducted a systematic review and meta-analysis of studies with prospective data  
90 collection in assessing the diagnostic accuracy of AI compared with trained HGs in screening for RDR  
91 in real-world settings. The findings of this review may offer evidence-based recommendations for  
92 integrating AI solutions to screen for RDR, especially in resource-challenged environments.

93

## 94 **Methods**

### 95 **Reporting, protocol, and registration**

96 We drafted this review in accordance with the Preferred Reporting Items for Systematic Review and  
97 Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) guidelines [12]. The study  
98 protocol was registered with the International Prospective Register of Systematic Reviews  
99 (PROSPERO) under CRD42023392297. An ethics waiver was granted by the University of Cape Town  
100 Human Research Ethics Committee.

## 101 **Databases and search strategies**

102 We searched the following electronic databases: Cochrane Central Register of Controlled Trials  
103 (CENTRAL), Medical Literature Analysis and Retrieval System Online (MEDLINE) via PubMed,  
104 Cumulative Index to Nursing and Allied Health Literature (CINAHL), Scopus, and Web of Science (**S1**  
105 **Table**). We also hand-searched the reference lists of relevant primary studies, systematic reviews, and  
106 the following journals: British Journal of Ophthalmology, American Journal of Ophthalmology,  
107 Ophthalmology and Retina, JAMA Ophthalmology, and Investigative Ophthalmology and Visual  
108 Science.

109

## 110 **Eligibility criteria**

### 111 **Type of studies**

112 We included randomised control trials (RCT) and observational analytical studies evaluating the DTA  
113 of AI in DR screening. We excluded studies based on retrospective validation of existing images (i.e.,  
114 medical records, available data sets). We excluded review articles, editorials, case series, case reports,  
115 and qualitative studies.

### 116 **Type of participants**

117 We included participants with clinically diagnosed type 1 or type 2 diabetes with unknown DR status,  
118 regardless of age, sex, race/ethnicity, and geographical location. We excluded studies that enrolled  
119 participants with unconfirmed diabetes to avoid misclassifying participants, which may result in biased  
120 estimates of the association between diabetes and diabetic retinopathy.

### 121 **Setting**

122 We only included studies conducted in real-world settings, thus excluding those done for theoretical  
123 algorithm training and validation alone.

124 **Index test**

125 We included interventions using AI for prospective screening of fundus images that could detect RDR  
126 or its equivalent.

127 **Reference standard**

128 The reference standard was manual grading for DR by trained HGs who analysed the same fundus  
129 images read by the AI. We excluded reference standards that did not use the same DR classification  
130 criteria used by the AI during its software training to grade DR.

131 **Target condition**

132 We included studies that screened for RDR as defined by the authors of the primary studies. We  
133 included studies with RDR equivalence, i.e. more than mild DR, clinically significant DR, etc. We did  
134 not include patients or eyes with no RDR, and ungradable or inconclusive fundus images in the pooling  
135 of diagnostic accuracy outcomes. Including ungradable or inconclusive images may result in inaccuracy  
136 in assessing the AI system's performance, making it challenging to draw meaningful conclusions.

137 **Outcomes**

138 We included studies reporting on, or containing the data necessary to extract information on the  
139 proportions of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).  
140 Efforts were made to contact corresponding authors to retrieve data which were unclear or unavailable  
141 in the paper or supplementary materials.

142 **Report characteristics**

143 We had no restrictions on the publication year and language. Study protocols were excluded.

144

145

146

147

## 148 **Study selection**

149 We used Rayyan software to manage the retrieved studies. Review authors (HU, CF) independently  
150 screened the titles and abstracts and classified them as (a) included, (b) maybe, and (c) excluded. Full-  
151 text articles of those ‘included’ and ‘maybe’ were obtained and independently assessed by the same  
152 authors against the eligibility criteria. Studies were then classified as (a) included, (b) excluded, and (c)  
153 awaiting authors’ responses. Any disagreements were resolved between the two reviewers or by  
154 consulting a third review author (AH). We emailed the corresponding authors of studies included as  
155 ‘awaiting authors’ responses’ at least three times with intervals of at least two weeks. If there were no  
156 responses from the authors, studies were classified under ‘no author’s response’.

## 157 **Data extraction and management**

158 We developed a data extraction form and divided it into two parts: (a) Study characteristics (relating to  
159 study designs, AI, and reference standards) and (b) Study outcomes: TP, FP, FN, TN. Two review  
160 authors extracted the study characteristics and study outcomes.

## 161 **Risk of bias and acceptability**

162 The risk of bias and applicability on the (a) patient selection, (b) index test, (c) reference standard, and  
163 (d) flow and timing of the included studies were independently assessed by two review authors (HU,  
164 CF) using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 tool [13]. We tailored,  
165 piloted, and refined our QUADAS-2 tool based on our review. Any disagreements were resolved  
166 between the two authors or by consulting a third review author (ME).

## 167 **Data synthesis and analysis**

### 168 **Quantitative data analysis and synthesis**

169 We calculated each included study’s sensitivity and specificity. We initially planned to analyse data  
170 only at the patient level; however, some studies reported only diagnostic accuracy on eye level (or image  
171 level), and some patient-level data cannot be extracted. Therefore, we considered looking into both of  
172 these levels for analysis. Heterogeneity was explored using visual inspection of forest plots and

173 hierarchical summary receiver operating characteristics (HSROC) plots. All analyses performed and  
174 plots generated were done using Review Manager (RevMan) 5.4 and SAS® Studio.

## 175 **Subgroup analysis**

176 We performed subgroup analyses on the following covariates identified a priori: level of economic  
177 development (World Bank country classification), level of the healthcare setting, and DR classification  
178 criteria. We did not include the modes of AI as previously planned since all AI modes of the included  
179 studies were automated.

## 180 **Sensitivity analysis**

181 We initially planned to explore the effect of excluding studies with a high risk of bias. However, after  
182 excluding studies with a high risk of bias, all studies were left with an unclear risk. Nevertheless, we  
183 performed sensitivity analyses to investigate the exclusion of studies that did not include DME in the  
184 RDR definition; although we have stated that the definition of RDR will be according to how the authors  
185 of the primary studies defined it, many references still included DME as part of RDR definition, and  
186 the International Council of Ophthalmology (ICO) guidelines states that patients with DME should be  
187 referred. We also investigated the exclusion of studies with  $\leq 2$  HGs as the ground truth for reference  
188 standard because this might incur bias if intergrader disagreements arise without having a third HG to  
189 arbitrate. According to Cardoso et al., ground truth means “data and/or method related to more  
190 consensus or reliable values/aspects that can be used as references” [14]. In our review, it refers to the  
191 final grading or assessment of fundus images by all HGs, which serves as the reference standard or the  
192 most reliable evaluation of the presence and severity of DR.

193

# 194 **Results**

## 195 **Results of the Search**

196 We were able to identify a total of 3899 articles through searching of various databases. After  
197 deduplication, 2742 studies were screened by title/abstract, of which 2654 were excluded. The  
198 remaining 88 studies were screened for full-text assessment against the review’s eligibility criteria. Of



199 these, 70 studies were excluded, three were classified under ‘no author’s response’, and finally, 15 were  
200 included for the quantitative synthesis (**Fig 1**).

201 **Fig 1. PRISMA flow diagram of the study search and selection.**

202

## 203 **Included studies**

204 Please see **Tables 1a** and **1b** for the characteristics of included studies. Fifteen studies comprising 17  
205 datasets were deemed eligible for this review, of which ten measured diagnostic accuracies at the patient  
206 level (45 785 patients) and seven at the eye level (15 390 eyes). We deemed all studies to be cross-  
207 sectional with prospective data collection; however, in our table of included studies, we presented the  
208 study designs according to how they were reported. Seven studies were done in China, five in India,  
209 and three in Australia. Seven studies were done at tertiary-level healthcare settings, six were done at  
210 the primary level, while the remaining two were done at both levels; no studies were done at the  
211 secondary level. For the target condition (RDR), six studies defined it as moderate NPDR or worse  
212 and/or DME, and nine did not include DME as part of the definition. Thirteen studies used ICDR or its  
213 equivalence as DR classification criteria, and two used the NHS DES criteria. For the reference  
214 standard, 12 studies have  $\geq 3$  HGs as the ground truth, and three studies have at most two HGs. Four  
215 studies developed their own AI models, and 11 used commercially available models. All studies used  
216 Inception with varying versions as their architecture. All AI software in the studies were fine-tuned with  
217 training data sets containing 25 297 to 207 228 fundus images. All studies used nonmydriatic cameras  
218 to capture fundus images, of which three still performed mydriasis on their patients using tropicamide  
219 eye drops, and one did mydriasis on a conditional protocol. Eight studies captured only one fundus field  
220 per eye (mostly macula-centred), and seven studies captured more than one fundus field. All studies  
221 used a fundus camera with a narrow field of vision (45°-50°).

**Table 1a. Key characteristics of the study design, population, target condition, and reference standard of included studies.**

Study	Study Settings				Patient Characteristics			Target Condition		Reference Standard/ Ground Truth (№)	
	Study Design <sup>a</sup>	Country (WBC)	Setting (№)	Healthcare Setting	Age, mean (SD), years	Type of Diabetes	Sample Size <sup>b</sup> (Patients/Eyes)	Definition of RDR/ Equivalence	Criteria Used	If without disagreement	If with disagreement
Dong 2022 [15]	Cross-sectional	China (U-MIC)	Community healthcare centres (3)	Primary	52.09 (±11.51)	T1D, T2D	Eyes: 848	Moderate NPDR or worse (DME not included)	ICDR	Gradings made by the ophthalmologists (2)	Gradings made by a senior retinal specialist (1)
Gulshan 2019 [16]	Prospective observational	India (L-MIC)	Eye care centre (Aravind Eye Hospital only)	Tertiary	56.60 (±9.00)	T1D, T2D	Eyes: 1905 <sup>c</sup>	Moderate NPDR or worse (DME not included)	ICDR	Gradings made by retinal specialists (3)	any disagreements discussed until a full consensus was achieved
Hao 2022 [17]	Prospective clinical trial	China (U-MIC)	Local community hospital	Primary	63.03 (±8.72)	T1D, T2D	Eyes: 6854	Moderate NPDR or worse (DME not included)	ICDR	Gradings made by the ophthalmologists (2)	Gradings made by a senior ophthalmologist (1)
He 2020 [18]	Cross-sectional <sup>d</sup>	China (U-MIC)	Community hospital clinic	Primary	68.46 (±7.20)	T1D, T2D	Patients: 889	Moderate NPDR or worse and/or DME	ICDR	Gradings made by the retina specialists (2)	Gradings made by a third retinal specialist (1)
Jain 2021 [19]	Cross-sectional	India (L-MIC)	Municipal dispensaries (47)	Primary	54.90 (±10.43)	T1D, T2D	Patients: 1370 Eyes: 2626	Moderate NPDR or worse (DME not included)	ICDR	Gradings made by the retina specialists (2)	Gradings made by a third retinal specialist (1)
Kanagas-ingam 2018 [20]	Cross-sectional <sup>d</sup>	Australia (HIC)	Primary care clinic	Primary	55.00 (±17.00)	T1D, T2D	Patients: 193	Moderate NPDR or worse (DME not included)	ICDR	Grading made by an ophthalmologist (1) alone	
Keel 2018 [21]	Prospective observational	Australia (HIC)	Urban endocrinology outpatient clinics (2)	Tertiary	44.26 (±16.56)	T1D, T2D	Patients: 93	Moderate NPDR or worse and/or DME	NHS DES	Grading made by the centralised retinal grading centre	
Li 2021 [22]	Prospective observational	China (U-MIC)	General hospital	Tertiary	50.00 (±12.00)	T1D, T2D	Eyes: 1674 <sup>c</sup>	Moderate NPDR or worse (DME not included)	ICDR	Grading made by a retina specialist (1) alone	
Natarajan 2019 [23]	Prospective, cross-sectional	India (L-MIC)	Municipal dispensaries	Primary	53.10 (±10.30)	T1D, T2D	Patients: 214 Eyes: 394	Moderate NPDR and worse, with or without DME	ICDR	Grading made by the ophthalmology resident (1) and retina specialist (1)	Gradings made by the same retina specialist
Rajalakshmi 2018 [24]	Cross-sectional <sup>d</sup>	India (L-MIC)	Diabetes centre	Tertiary	NR	T2D	Patients: 296	Moderate NPDR or worse and/or DME	ICDR	Gradings made by the retina specialists (2)	Gradings made by a third retinal specialist (1)
Scheetz 2021 [25]	Prospective observational	Australia (HIC)	Endocrinology outpatient clinics (2) and Aboriginal medical services clinics (3)	Primary and Tertiary	54.25 (±20.16) <sup>e</sup>	T1D, T2D	Patients: 203	Moderate NPDR or worse and/or DME	NHS DES	Gradings made by NHS-certified graders (2)	Gradings made by retinal specialists (2)
Sosale 2020 [26]	Prospective, cross-sectional	India (L-MIC)	Diabetes centre	Tertiary	NR	T1D, T2D	Patients: 900	Moderate NPDR or worse and/or DME	ICDR	The majority diagnosis of the retina specialists (5)	
Yang 2022 [27]	Observational, prospective, multicentre, gold standard-controlled	China (U-MIC)	Hospital and ophthalmic centres (3)	Tertiary	60.44 (±10.19) <sup>e</sup>	T1D, T2D	Patients: 962	Stage II or worse DR (DME not included)	COS <sup>f</sup>	Gradings made by ZIRC graders (2)	Gradings made by a third senior ZIRC grader (1)
Zhang 2020 [28]	Prospective observational	China (U-MIC)	Diabetes centres (155)	Primary and Tertiary	54.29 (±11.60)	T1D, T2D	Patients: 40 665	Moderate NPDR or worse (DME not included)	ICDR	Gradings made by the ophthalmologists (2)	Gradings made by a senior ophthalmologist (1)
Zhang 2022 [29]	Prospective, multicentre, self-controlled clinical trial	China (U-MIC)	Hospitals (3)	Tertiary	56.52 (±11.13)	T1D, T2D	Eyes: 1089	Moderate NPD or worse (DME not included)	ICDR	Gradings made by the ophthalmologists (3)	Gradings made by the principal investigator ophthalmologist (1)

<sup>a</sup> Study design according to study authors; <sup>b</sup> Sample included in the diagnostic accuracy analysis excluding ungradable images; <sup>c</sup> Samples were reported in image level, but the study captured one image per eye, so considered as eye-level;

<sup>d</sup> Study design not reported, thus deemed by review authors as cross-sectional based on the journals; <sup>e</sup> Mean was estimated from median using recommendations by Hong Kong Baptist University, Department of Mathematics [30];

<sup>f</sup> Criteria was matched to the equivalent definition of RDR based on the ICDR classification.

**COS**, Chinese Ophthalmic Society; **CSME**, clinically significant macular edema; **DME**, diabetic macular edema; **DR**, diabetic retinopathy; **ETDRS**, Early Treatment Diabetic Retinopathy Study; **HIC**, high-income country; **ICDR**, International Clinical Diabetic Retinopathy; **L-MIC**, lower middle-income country; **NHS DES**, National Health Service Diabetic Eye Screening; **NPDR**, nonproliferative diabetic retinopathy; **NR**, not reported; **RDR**, referable diabetic retinopathy; **SD**, standard deviation; **T1D**, Type 1 diabetes; **T2D**, Type 2 diabetes; **U-MIC**, upper middle-income country; **WBC**, World Bank classification; **ZIRC**, Zhongshan Image Reading Centre.

**Table 1b. Key characteristics of the index tests of included studies.**

Study	Artificial Intelligence Development						Fundus Camera Used		
	AI Model	Architecture	Neural Network	Pre-trained	Fine-tuned	Training Dataset (№ of fundus images)	Mydriatic or Nonmydriatic Camera	№ of Fundus Fields	Field of Vision
Dong 2022 [15]	CARE, Shanghai EagleVision Medical Technology Co., Ltd (Airdoc)	Inception-ResNet-v2	CNN	Yes	Yes	Clinical settings datasets (207 228)	Nonmydriatic	1 field (macula-centred)	50°
Gulshan 2019 [16]	Own AI model	Inception-v3	CNN	Yes	Yes	EyePACS and hospital datasets (128 175)	Nonmydriatic	1 field (macula-centred)	45°
Hao 2022 [17]	EyeWisdom (Visionary Intelligence Ltd., Beijing, China)	Inception-v3	CNN	Yes	Yes	EyePACS and hospital datasets (25 297)	Nonmydriatic	2 fields (macula- and optic disc-centred)	45°
He 2020 [18]	Airdoc, Beijing, China	Inception-v4	CNN	Yes	Yes	Unspecified dataset (number of fundus images NR)	Nonmydriatic	2 fields (macula- and optic disc-centred)	45°
Jain 2021 [19]	Medios AI (Remidio)	Inception-v3 and MobileNet	CNN	Yes	Yes	EyePACS, hospital and screening camps datasets (52 894)	Nonmydriatic, but patients underwent mydriasis (1% tropicamide)	3 fields (posterior pole including macula & disc, nasal, and temporal)	45°
Kanagas-ingam 2018 [20]	Own AI model	Inception-v3 (customised)	CNN	Yes	Yes	DiaRetDB1, EyePACS, and Tele-eye care DR database (30 000)	Nonmydriatic	1 field (macula-centred)	45°
Keel 2018 [21]	Own AI model	Inception-v3	CNN	NR	Yes	LabelMe dataset (58 790)	Nonmydriatic	1 field (central nasal)	45°
Li 2021 [22]	VoxelCloud, China	Inception-ResNet-v2	CNN	Yes	Yes	EyePACS and hospital datasets (141 184)	Nonmydriatic	1 field (macula-centred)	45°
Natarajan 2019 [23]	Medios AI (Remidio)	Inception-v3 and MobileNet	CNN	Yes	Yes	EyePACS, hospital and screening camps datasets (52 894)	Nonmydriatic, but patients underwent mydriasis (1% tropicamide)	3 fields (posterior pole including macula & disc, nasal, and temporal)	45°
Rajalakshmi 2018 [24]	EyeArt v2.1	NR	DNN	Yes	Yes	EyePACS (number of fundus images NR)	Nonmydriatic, but patients underwent mydriasis (tropicamide)	4 fields (macula-centred, optic disc-centred, superior-temporal, and inferior-temporal quadrants of the retina)	45°
Scheetz 2021 [25]	Own AI model	Inception-v3	CNN	NR	Yes	LabelMe dataset (71 043)	Nonmydriatic	1 field (macula-centred)	45°
Sosale 2020 [26]	Medios AI (Remidio)	Inception-v3 and MobileNet	CNN	Yes	Yes	EyePACS, hospital and screening camps datasets (52 894)	Nonmydriatic	2 fields (macula- and optic disc-centred)	45°
Yang 2022 [27]	AIDRScreening v1.0 (Shenzhen SiBright CO. Ltd., China)	NR	CNN	NR	Yes	Eye institute, endocrinology department, and eye examination centre datasets (73 849)	Both; if pupil diameter was >4 mm, fundus photography was performed without mydriasis; otherwise, mydriasis was required	2 fields (macula- and optic disc-centred)	45°
Zhang 2020 [28]	VoxelCloud Retina, China	Inception-ResNet-v2	CNN	Yes	Yes	EyePACS and hospital datasets (144 810)	Nonmydriatic	1 field (macula-centred)	45°
Zhang 2022 [29]	EyeWisdom v1 (Visionary Intelligence Ltd., Beijing, China)	Inception-v3 and ResNet-34	CNN	Yes	Yes	Hospital and ILSVRC subset of ImageNet datasets (40 693)	Nonmydriatic	1 field (posterior pole containing macula and optic disc)	45°

CARE, Comprehensive AI Retinal Expert; CNN, convolutional neural network; DNN, deep neural network; ILSVRC, ImageNet Large Scale Visual Recognition Challenge; NR, not reported

## 223 **Excluded studies**

224 From the 88 full-text articles assessed for eligibility, we excluded 70 studies and classified three  
225 studies under ‘no author’s response’.

## 226 **Methodological quality of included studies**

227 A summary of methodological quality assessment is presented in **Figs 2** and **3**.

228 **Fig 2. Risk of bias and applicability concerns summary: Review authors' judgments about each**  
229 **domain for each included study using the QUADAS-2 tool.**

230 **Fig 3. Risk of bias and applicability concerns graph: Review authors' judgments about each**  
231 **domain presented as percentages across included studies using the QUADAS-2 tool.**

232

## 233 **Patient selection**

234 In the patient selection domain, 12 of the 15 studies were deemed to have an unclear risk of bias in the  
235 sampling method. Most of the studies did not specify how patients were enrolled except for three studies  
236 (1 consecutive, 1 random, and 1 convenience sampling method). Two studies were not able to avoid  
237 inappropriate exclusions since one study excluded patients with macular edema, and the other study  
238 excluded those who were treated with ocular injections for DME or proliferative disease; of which these  
239 conditions are part of the definition of RDR, deeming these studies with a high risk of bias and high  
240 concern on applicability. For applicability on patient selection, 13 out of 15 studies have a low concern  
241 on applicability.

242

## 243 **Index test**

244 In the domain of index tests, we added two signalling questions deemed necessary for index tests using  
245 AI, one of which is the quality of images fed into the AI system. This is vital since images with  
246 insufficient quality (i.e., overexposed, out-of-focus, etc.) may be deemed ungradable or be  
247 misclassified. Another signalling question added was on the conflict of interest. With the advent of AI

248 in healthcare, several AI software packages are currently being developed; thus, if study authors were  
249 affiliated with or funded by the software company in any way, studies may incur a high risk of bias.

250 In this domain, the main quality issue was the signalling question of whether a diagnostic threshold was  
251 prespecified or not. Only three studies reported on prespecified thresholds, with the remaining 12  
252 studies thus considered to have an unclear risk of bias. Six studies have conflicts of interest, thus  
253 deeming them high risk. All studies have low applicability concerns for the index test.

254

## 255 **Reference standard**

256 In the reference standard domain, three out of 15 studies were evaluated as having a high risk of bias  
257 because there were only two HGs to grade the fundus images. This may incur bias since grading images  
258 can be very subjective, and there is no one to arbitrate when a disagreement arises. Five studies have an  
259 unclear risk of bias because they did not explicitly state whether the HGs were blinded to the results of  
260 the AI grading results. All studies have low applicability concerns.

261

## 262 **Flow and timing**

263 In the domain of flow and timing, one study was considered to be of a high risk of bias because it was  
264 not able to explain the discrepancies in patients enrolled and analysed clearly. This domain is not  
265 assessed regarding applicability concerns, as stated in the QUADAS-2 tool.

266

## 267 **Findings**

268 We evaluated the accuracy of AI in screening for RDR in real-world settings according to patient-level  
269 and eye-level analysis compared with HGs. The patient-level analysis was considered the main meta-  
270 analysis since it is the number of patients with RDR who will be referred to ophthalmologists for further  
271 assessment. Out of the 15 studies reviewed, eight presented diagnostic accuracy based solely on patient-

272 level information, five showed diagnostic accuracy based solely on eye-level information, and two  
273 showed diagnostic accuracy based on both patient-level and eye-level information.

274 The HSROC model by Rutter and Gatsonis was used for the meta-analysis as this model accounts for  
275 the variations in the test thresholds among the AI models [31]. We performed subgroup analysis and  
276 investigated for heterogeneity using the World Bank country classification, level of the healthcare  
277 setting, and DR classification criteria.

278 We performed sensitivity analyses to explore the effect of excluding (1) studies that did not include  
279 DME in the RDR definition and (2) studies with a total number of  $\leq 2$  HGs as the ground truth. **Table**  
280 **2** shows the detailed overall patient-level and eye-level meta-analysis.

281

**Table 2. Overall patient-level and eye-level meta-analysis of the accuracy of AI in detecting RDR compared with trained HGs.**

Overall Meta-analysis	N <sup>o</sup> of Studies	N <sup>o</sup> of Samples	Sensitivity (95% CI)	Specificity (95% CI)
Patient-level	10	45 785 patients	95.33% (90.60-100)	92.01% (87.61-96.42)
Eye-level	7	15 390 eyes	91.24% (79.15-100)	93.90% (90.63-97.16)

Data calculated using SAS<sup>®</sup> Studio.

CI, Confidence Interval; HG, human grader.

282

### 283 **Patient-level analysis**

284 Ten evaluations of AI for RDR screening were performed with data from ten studies and a total of 45  
285 785 patients. The forest plot (**Fig 4**) shows minimal variation in the accuracy estimates. The HSROC  
286 plot (**Fig 5**) reveals good test accuracy since most study points lie in the upper left corner of the plot.  
287 Meta-analytical sensitivity and specificity of data at mixed thresholds were 95.33% (95% CI 90.60-  
288 100) and 92.01% (95% CI 87.61-96.42), respectively.

289

290 **Fig 4. Coupled forest plot of included studies for patient-level analysis.**

291 **Fig 5. HSROC plot of sensitivity vs specificity of AI for detecting RDR on patient-level analysis.**

292

## 293 **Eye-level analysis**

294 A total of seven evaluations of AI for RDR screening were performed with data from seven studies and  
295 a total of 15 390 eyes. We only included the Aravind data from the Gulshan 2019 study because data  
296 from Sankara differed from our eligibility criteria.

297 The forest plot (**Fig 6**) shows moderate variation in the estimates of sensitivity and minimal variation  
298 in specificity. The HSROC plot (**Fig 7**) reveals good test accuracy since most study points lie in the  
299 upper left corner of the plot. Meta-analytical sensitivity and specificity of data at mixed thresholds were  
300 91.24% (95% CI 79.15-100) and 93.90% (95% CI 90.63-97.16), respectively.

301

302 **Fig 6. Coupled forest plot of included studies for eye-level analysis.**

303 **Fig 7. HSROC plot of sensitivity vs specificity of AI for detecting RDR on eye-level analysis.**

304

## 305 **Exploring heterogeneity**

306 We performed subgroup analyses to explore potential sources of heterogeneity only on the main  
307 analysis (patient level), consisting of ten studies, since the data for the subgroups were more complete.

308 A detailed result of subgroup analyses investigating potential sources of study-level heterogeneity is  
309 shown in **Table 3**.

310

311

312

313

314

315

316

**Table 3. Subgroup analyses for the accuracy of AI in detecting RDR compared with trained HGs on patient-level analysis.**

Analysis		N <sup>o</sup> of Studies	N <sup>o</sup> of Participants	Sensitivity (95% CI)	Specificity (95% CI)
<b>Overall Meta-analysis</b>					
Patient-level		10	45 785	95.33% (90.60-100)	92.01% (87.61-96.42)
<b>Subgroup Analyses</b>					
World Bank Country Classification	LMIC	7	45 296	95.38% (90.38-100)	92.21% (87.19-97.23)
	HIC	3	489	95.61% (89.44-100)	90.82% (87.76-93.87)
Level of Health-care Setting <sup>a</sup>	Primary	4	2666	99.35% (96.85-100)	93.72% (88.83-98.61)
	Tertiary	4	2251	94.71% (89.00-100)	90.88% (83.22-98.53)
DR Classification Criteria	ICDR	8	45 489	95.44% (90.70-100)	92.21% (87.80-96.62)
	NHS DES	2	296	95.49% (89.19-100)	89.85% (84.93-94.77)

<sup>a</sup> No studies reported on secondary healthcare settings, thus not included in this table.

Data calculated using SAS<sup>®</sup> Studio.

**CI**, Confidence Interval; **HG**, human grader; **HIC**, high-income country; **ICDR**, International Clinical Diabetic Retinopathy; **LMIC**, lower- and middle-income country; **NHS DES**, National Health Service Diabetic Eye Screening; **RDR**, referable diabetic retinopathy.

317

### 318 **Level of economic development**

319 We classified the level of economic development of the countries included in our study using  
 320 classification by the World Bank Group [32]. Of the ten studies included, three were conducted in HICs,  
 321 and seven were conducted in LMICs. Australia was classified as a high-income country (HIC), and  
 322 China and India, as lower- and middle-income country (LMIC). The sensitivity and specificity of AI in  
 323 the real-world screening for RDR in LMICs were 95.38% and 92.21%, respectively, and in HIC, they  
 324 were 95.61% and 90.82%, respectively (**Fig 8**).

325

326 **Fig 8. Coupled forest plots showing the subgroups in the level of economic development**  
 327 **according to the World Bank country classification.**

328 **HIC**, high-income country; **LMIC**, lower- and middle-income country

329

330



331 **Level of healthcare setting**

332 Four studies were done solely in tertiary-level healthcare settings, four in primary-level healthcare  
333 settings, and two at both levels (which were not included in this analysis). The sensitivity and specificity  
334 of AI in the real-world screening for RDR in primary-level healthcare settings were slightly higher than  
335 in tertiary-level (99.35% vs 94.71%, and 93.72% vs 90.88%, respectively) (**Fig 9**).

336

337 **Fig 9. Coupled forest plots showing the subgroups in the level of healthcare settings.**

338

339 **DR classification criteria**

340 Eight studies used ICDR or its equivalence as DR classification criteria, and only two used the NHS  
341 DES criteria. It is important to note that doing a subgroup in this covariate does not intend to compare  
342 the two criteria but rather to see the robustness of AI in screening for RDR, even using different criteria.  
343 The sensitivity and specificity of AI in the real-world screening for RDR using ICDR were 95.45% and  
344 92.21%, respectively, and using NHS DES, they were 95.49% and 89.85%, respectively, which did not  
345 show any significant variation (**Fig 10**).

346

347 **Fig 10. Coupled forest plots showing the subgroups in the DR classification criteria.**

348 **ICDR**, International Clinical Diabetic Retinopathy; **NHS DES**, National Health Service Diabetic Eye  
349 Screening.

350

351 **Sensitivity analysis**

352 We performed sensitivity analyses on two conditions stated below. A detailed result of sensitivity  
353 analyses is shown in **Table 4**.

354

355

356

**Table 4. Sensitivity analyses for the accuracy of AI in detecting RDR compared with trained HGs on patient-level analysis.**

Analysis	N <sup>o</sup> of Studies	N <sup>o</sup> of Participants	Sensitivity (95% CI)	Specificity (95% CI)
<b>Overall Meta-analysis</b>				
Patient-level	10	45 785	95.33% (90.60-100)	92.01% (87.61-96.42)
<b>Sensitivity Analyses</b>				
Only studies with DME included in the RDR definition	6	2595	95.51% (92.58-98.44)	91.35% (84.92-97.78)
Only studies with a total of $\geq 3$ HGs	8	45 378	94.69% (90.11-99.28)	92.37% (87.93-96.81)

Data calculated using SAS<sup>®</sup> Studio.

CI, confidence interval; DME, diabetic macular edema; HG, human grader; RDR, referable diabetic retinopathy.

357

### 358 **Inclusion of DME on the RDR definition**

359 After excluding four studies that did not include DME as part of the RDR definition, pooled sensitivity  
 360 and specificity did not show any significant variation compared to the overall main meta-analysis  
 361 (95.51% vs 95.33% and 91.35% vs 92.01%, respectively) (**Fig 11**).

362

### 363 **Fig 11. Coupled forest plot of studies that include DME on the RDR definition.**

364 DME, diabetic macular edema; RDR, referable diabetic retinopathy.

365

### 366 **Total number of human graders**

367 After excluding two studies that have a total of  $\leq 2$  HGs as the ground truth, pooled sensitivity and  
 368 specificity also did not show any significant variation compared to the overall main meta-analysis  
 369 (94.69% vs 95.33% and 92.37% vs 92.01%, respectively) (**Fig 12**).

370

### 371 **Fig 12. Coupled forest plot of studies with $\geq 3$ human graders as the ground truth on reference** 372 **standard.**

373

## 374 **Investigation of publication bias**

375 We did not investigate publication bias since, according to Salameh, et al. [12], the statistical  
376 investigation of publication and reporting bias is not routinely recommended in systematic reviews  
377 involving DTA.

378

## 379 **Discussion**

### 380 **Summary of main findings**

381 Artificial intelligence screening incorporating a range of software applications has been evaluated for  
382 detecting referable DR in real-world settings. Studies in this review came from various economic  
383 settings and level of health care, all using recognised DR classification criteria. This review provides  
384 evidence, for the first time from prospective studies, for the effectiveness of AI in screening for RDR,  
385 in real-world settings.

386 This review aimed to assess the accuracy of AI solutions in detecting RDR in different resource settings.  
387 We found no variation in the diagnostic accuracy of AI, whether deployed in the LMICs or HICs,  
388 meaning AI in screening for RDR can be used universally. Regarding different DR classifications used,  
389 we found no variation between ICDR and NHS DES because both the AI models and HGs used the  
390 same criteria when grading RDR. Thus, stakeholders need to note that when integrating an AI model  
391 into a DR screening programme, the DR criteria used to train the AI model should be the same as the  
392 DR criteria used by the trained HGs in that setting or country to prevent misclassifications.

393 However, on the level of the healthcare setting, studies done in the primary-level healthcare settings  
394 have higher diagnostic accuracy compared to those done in tertiary-level healthcare settings. One of the  
395 reasons may be having more patients with advanced disease or other comorbidities in tertiary care  
396 settings where screening for RDR can be more challenging.

397 We applied the summary estimates to a hypothetical cohort of 1000 patients to our main analysis using  
398 the Grading of Recommendations, Assessment, Development and Evaluation (GRADE)pro guideline  
399 development tool [33] (**Table 5**). Our findings suggest that if AI is used for the detection of RDR in

400 real-world settings, 95% of patients with RDR will be correctly screened positive for the condition, and  
401 92% of patients with no RDR will be correctly screened negative for the condition. We were interested  
402 in knowing the number of patients who will be correctly and unnecessarily referred to tertiary healthcare  
403 for a further eye examination. For our prevalence rate, we used a prevalence estimate of 6.5%, which is  
404 from a recent multi-ethnic study involving datasets from Singapore, the USA, China and Australia [34]  
405 and a prevalence estimate of 2%, which is the national prevalence estimate of RDR in India [35]. Using  
406 an RDR prevalence of 6.5%, AI will correctly detect RDR in 62 patients living with diabetes, miss  
407 detecting three RDR cases while unnecessarily refer 75 patients living with diabetes without RDR, for  
408 further examination.  
409

410

**Table 5. Summary of findings of the review evaluated using the GRADEpro GDT.**

<b>Review question:</b> What is the diagnostic test accuracy of AI in screening for RDR compared with trained HGs among patients with diabetes in real-world settings?					
<b>Population:</b> People living with clinically diagnosed type 1 and type 2 diabetes					
<b>Setting:</b> Real-world settings					
<b>Index test:</b> Artificial intelligence					
<b>Reference standard:</b> Trained HGs					
<b>Study design:</b> Cross-sectional studies with prospective data collection					
<b>Total № of studies:</b> 15 studies; <b>Patient-level (Main) analysis:</b> 10 studies (45 785 patients); <b>Eye-level analysis:</b> 7 studies (15 390 eyes)					
Effect (95% CI)	Test Result	№ of results per 1000 Samples Tested (95% CI)		№ of Samples (Studies)	Certainty of the Evidence (GRADE)
		Prevalence 2% <sup>a</sup>	Prevalence 6.5% <sup>b</sup>		
<b>Patient-level analysis</b>					
Pooled sensitivity 95% (91-100%)	True Positive	19 (18-20)	62 (59-65)	10 985 patients (10 studies)	⊕⊕⊕○ <b>MODERATE</b> <sup>c</sup>
	False Negative	1 (0-2)	3 (0-6)		
Pooled specificity 92% (88-96%)	True Negative	902 (859-945)	860 (819-902)	34 890 patients (10 studies)	⊕⊕⊕○ <b>MODERATE</b> <sup>c</sup>
	False Positive	78 (35-121)	75 (33-116)		
<b>Eye-level analysis</b>					
Pooled sensitivity 91% (79-100%)	True Positive	18 (16-20)	59 (51-65)	2913 eyes (7 studies)	⊕○○○ <b>VERY LOW</b> <sup>d, e, f</sup>
	False Negative	2 (0-4)	6 (0-14)		
Pooled specificity 94% (91-97%)	True Negative	920 (888-952)	878 (847-908)	12 477 eyes (7 studies)	⊕⊕⊕○ <b>MODERATE</b> <sup>d</sup>
	False Positive	60 (28-92)	57 (27-88)		
Prevalence data calculated using GRADEpro GDT. <sup>a</sup> National prevalence estimate of RDR in India [35] <sup>b</sup> Prevalence estimate of RDR in a multi-ethnic study involving datasets from Singapore, the USA, Hong Kong, China and Australia [34] <sup>c</sup> <b>Risk of bias (-1):</b> QUADAS-2 tool was used to assess for the risk of bias in the 10 studies. In the domain of <i>Patient Selection</i> , the risk of bias was high in 1 study and was unclear in 8; In the domain of <i>Index Test</i> , it was high in 4 studies and unclear in 6; In the domain of <i>Reference Standard</i> , it was high in 2 studies and unclear in 4; and in the domain of <i>Flow and Timing</i> , it was high in 1 study. <sup>d</sup> <b>Risk of bias (-1):</b> Risk of bias was assessed in the 7 studies of this level of analysis. In the domain of <i>Patient Selection</i> , the risk of bias was high in 2 studies and was unclear in 5; In the domain of <i>Index Test</i> , it was high in 2 studies and unclear in 4; In the domain of <i>Reference Standard</i> , it was high in 2 studies and unclear in 1; and in the domain of <i>Flow and Timing</i> , it was high in 1 study. <sup>e</sup> <b>Inconsistency(-1):</b> Statistical heterogeneity based on the forest plot showed moderate variation in the sensitivity. <sup>f</sup> <b>Imprecision (-1):</b> The CI of the pooled sensitivity is wide, indicating that there is an uncertainty in the estimate and that the true value could potentially be lower. <b>Grade Definition</b> [33] <b>High:</b> Further research is very unlikely to change our confidence in the estimate of effect; <b>Moderate:</b> Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate; <b>Low:</b> Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate; <b>Very low:</b> Any estimate of effect is very uncertain. <b>AI,</b> artificial intelligence; <b>CI,</b> Confidence Interval; <b>DME,</b> diabetic macular edema; <b>DR,</b> diabetic retinopathy; <b>GDT,</b> guideline development tool; <b>GRADE,</b> Grading of Recommendations, Assessment, Development and Evaluation; <b>HG,</b> human grader; <b>RDR,</b> referable diabetic retinopathy					

411

412

413 We also explored the effect of excluding studies that did not include DME in the RDR definition and  
414 found that the diagnostic accuracy of AI has no significant variation; this does not mean that inclusion  
415 or exclusion of DME is nonsignificant in screening for RDR, rather, this is because trained HGs adhered  
416 to the grading protocol of the study with regards to RDR definition. Furthermore, the effect of excluding  
417 studies with  $\leq 2$  trained HGs (ophthalmologists and trained and certified HGs from retina reading  
418 centres) did not affect the diagnostic accuracy of the data. This review, thus, highlights the importance  
419 of trained HGs acting as a reference standard for grading the fundus images.

420

## 421 **Strengths and limitations of this review**

### 422 **Strengths**

423 This is the first systematic review and meta-analysis assessing the diagnostic accuracy of AI in  
424 screening for RDR in real-world settings that included studies using prospective data collection. We  
425 did not restrict our literature search in terms of language and publication year to minimise the chance  
426 of missing studies. We were able to present the accuracy estimates in patient-level and eye-level  
427 analysis, rather than just combining these data to prevent unit-of-analysis issues and avoid bias in  
428 precision. We were able to tailor and pilot our QUADAS-2 tool to our study, adding more signalling  
429 questions to fit AI studies since QUADAS-AI by Sounderajah et al. [36] was not yet published during  
430 the time of our review. Data extraction and assessment of the risk of bias were performed by two review  
431 authors, thus, reducing the risk of bias. We avoided all case-control studies since studies involving a  
432 control group without RDR and patients with RDR may exaggerate the diagnosis accuracy [13]. We  
433 included studies using different DR criteria (rather than just restricting to certain criteria), where results  
434 showed no significant variation in the accuracy estimates.

435

436

437

438

## 439 **Limitations**

### 440 **Eligibility**

441 Our definition of RDR in this review is according to how the authors of the primary studies defined  
442 them, with or without DME. In the clinical setting, cases of DME should be referred for further  
443 examination when detected. However, to accurately detect DME, optical coherence tomography (OCT),  
444 which gives detailed 3D images of the eye, is the gold standard; thus, this makes fundus images less  
445 advantageous as it only provides 2D images. Therefore, it is important that further studies be done for  
446 AI models to be trained and developed to read OCT together with fundus images for higher accuracy  
447 and better applicability.

448

### 449 **Quality of included studies**

450 All eligible studies had either an unclear risk or, a high risk of bias in at least one of the QUADAS-2  
451 domains. Amongst the included studies, 80% did not report as to how patients were enrolled in the  
452 study, making them unclear. Also, many of the studies did not clearly report a pre-specified threshold  
453 which may influence the diagnostic accuracy of the test if the authors select a positivity cut-off after  
454 obtaining the results. Thus, we support that DTA studies following the Standards for Reporting of  
455 Diagnostic Accuracy Studies (STARD) guidelines by Cohen et al. [37] or the proposed STARD-AI  
456 guidelines by Sounderajah et al. [38], when available, to avoid these uncertainties.

457 Regarding the QUADAS-2 domain on flow and timing, specifically as regards the signalling question  
458 relating to whether all enrolled patients were included in the analysis, we deemed a study as high risk  
459 if the discrepancies between the enrolled and analysed patients were not motivated, or were related to  
460 the severity of RDR (even though most studies have excluded ungradable images from the analysis).  
461 This was done since including ungradable images may lead to inaccuracy and not give meaningful  
462 results. Therefore, it is important that during the implementation of AI in DR screening programmes,  
463 the protocol for evaluating images as ungradable should be available, (e.g. considering mydriasis, if  
464 needed, assuring quality images when capturing photos, etc.), to avoid missed detections and

465 unnecessary referrals since during DR screening, patients with fundus images deemed ungradable by  
466 AI should also be referred to ophthalmologists for proper assessment.

467 Another limitation found is the representativity of the level of economic development by World Bank  
468 country classification. The subgroup for HIC is represented only by Australia, and the subgroup LMIC,  
469 only by China and India. Although there were DTA studies conducted in other countries (i.e. USA,  
470 Spain, Zambia, etc), they were, unfortunately, excluded against our eligibility criteria.

471

## 472 **Applicability of findings to the review question**

473 Concerns regarding the applicability of all included studies were deemed low, except for two studies  
474 that were not able to avoid inappropriate exclusions. We assessed the applicability of findings to our  
475 review question with low concerns since all studies included AI models that were able to detect RDR  
476 in real-world settings; included patients were all clinically diagnosed with type 1 and/or type 2 diabetes;  
477 the grading of the same images was all compared to the grading of the trained HGs.

478

## 479 **Conclusion**

480 Our review provides evidence that AI could effectively screen for RDR even in real-world settings.  
481 Whether in the HICs or LMICs, the detection of RDR using AI in real-world settings is highly sensitive  
482 and specific. It has higher accuracy when deployed at the primary-level than in tertiary-level healthcare  
483 settings.

484

## 485 **Implications for practice**

486 Although AI in screening for DR has been showing promising results, it is important to consider where  
487 to deploy them. Patient-wise, it will be able to screen more patients living with diabetes, leading to early  
488 diagnosis and treatment. It can also increase disease awareness, promoting a healthy lifestyle and  
489 diabetes control to these patients. However, healthcare-wise, AI might be unnecessarily referring a



490 handful of patients without RDR to tertiary healthcare centres. In HICs, where manpower is usually not  
491 an issue, this might not be a problem; however, in LMICs, where it is a challenge, referring false positive  
492 cases to the already few and straining eye health workers can overburden them. Thus, we recommend  
493 a clinical pathway in these low-resource settings, where trained or certified lay graders in primary  
494 healthcare can countercheck all the fundus images of patients who screened positive for RDR before  
495 officially referring them, rather than just leaving the referral decisions to the AI system.

496

## 497 **Implications for research**

498 In recent years, researchers and clinicians have been advocating the use of real-world performance of  
499 AI for healthcare to evaluate further their real impact on image quality and system usability rather than  
500 just validating them using retrospective high-quality databases [24]. Our review was able to pool the  
501 diagnostic accuracy of AI in screening RDR of studies using prospective data collection; therefore, can  
502 provide recommendations to evidence-based guidelines to integrate AI in DR screening programmes in  
503 real-world settings. We recommend further studies on integrating OCT aside from using fundus imaging  
504 in AI algorithms so screening for DME will be more accurate.

505

## 506 **Acknowledgements**

507 M.E.E. is supported by a grant from the South African Medical Research Council. The views and  
508 opinions expressed are those of the authors and do not necessarily represent the official views of the SA  
509 MRC.

510

## 511 **Author contributions**

512 **Conception of the review:** Holijah Uy, Deon Minnies, Abraham Opare

513 **Development of the protocol:** Holijah Uy, Deon Minnies, Ameer Hohlfeld, Mark E Engel

514 **Methodological advice:** Mark E Engel, Ameer Hohlfeld, Eleanor Ochodo

515 **Content advice:** Mark E Engel, Christopher Fielding, Deon Minnies

516 **Data collection:** Holijah Uy, Christopher Fielding, Elton Mukonda,

517 **Data analysis:** Holijah Uy, Eleanor Ochodo, Ameer Hohlfeld, Mark E Engel

518 **Writing of first draft:** Holijah Uy

519 **Contributions to editing subsequent versions of the protocol and manuscript:**

520 Ameer Hohlfeld, Eleanor Ochodo, Deon Minnies, Mark E Engel

521

## 522 **References**

523 1. Teo ZL, Tham Y-C, Yan Yu MC, Chee ML, Rim TH, Cheung N, et al. Global prevalence of  
524 diabetic retinopathy and projection of burden through 2045: Systematic review and meta-  
525 analysis. *American Academy of Ophthalmology*. 2021;128.

526 <https://doi.org/10.1016/j.ophtha.2021.04.027>

527 2. Flaxel CJ, Adelman RA, Bailey ST, Fawzi A, Lim JJ, Vemulakonda GA, et al. Diabetic  
528 retinopathy preferred practice pattern. *American Academy of Ophthalmology*. 2019;127: P66–  
529 P145. <https://doi.org/10.1016/j.ophtha.2019.09.025>

530 3. Cuadros J. The real-world impact of artificial intelligence on diabetic retinopathy screening in  
531 primary care. *Journal of Diabetes Science and Technology*. 2021;15: 664–665.

532 <https://doi.org/10.1177/1932296820914287>

533 4. International Council of Ophthalmology. Guidelines for diabetic eye care. San Francisco,  
534 California: ICO; 2017. Available: <https://icoph.org/eye-care-delivery/diabetic-eye-care/>

535 5. Wintergerst MWM, Bejan V, Hartmann V, Schnorrenberg M, Bleckwenn M, Weckbecker K, et  
536 al. Telemedical diabetic retinopathy screening in a primary care setting: Quality of retinal  
537 photographs and accuracy of automated image analysis. *Ophthalmic Epidemiology*. 2021;29:

538 286–295. <https://doi.org/10.1080/09286586.2021.1939886>

539

- 540       **6.** Piyasena MMPN, Murthy GVS, Yip JLY, Gilbert C, Zuurmond M, Peto T, et al. Systematic  
541       review on barriers and enablers for access to diabetic retinopathy screening services in different  
542       income settings. PLoS ONE. 2019;14. <https://doi.org/10.1371/journal.pone.0198979>
- 543       **7.** National Institute for Clinical Excellence. Clinical Guideline E: Management of Type 2  
544       diabetes. London: NICE; 2002.
- 545       **8.** London School of Hygiene & Tropical Medicine. Identifying a good screening test: sensitivity,  
546       specificity and coverage. Future Learn. Available:  
547       <https://www.futurelearn.com/info/courses/diabetic-eye-disease/0/steps/47640>
- 548       **9.** Silpa-archa S, Limwattanayingyong J, Tadarati M, Amphornphruet A, Ruamviboonsuk P.  
549       Capacity building in screening and treatment of diabetic retinopathy in Asia-Pacific region.  
550       Indian Journal of Ophthalmology. 2021;69: 2959–2967.  
551       [https://doi.org/10.4103/ijo.IJO\\_1075\\_21](https://doi.org/10.4103/ijo.IJO_1075_21)
- 552       **10.** Raman R, Dasgupta D, Ramasamy K, George R, Mohan V, Ting D. Using artificial intelligence  
553       for diabetic retinopathy screening: Policy implications. Indian Journal of Ophthalmology.  
554       2021;69: 2993–2998. [https://doi.org/10.4103/ijo.IJO\\_1420\\_21](https://doi.org/10.4103/ijo.IJO_1420_21)
- 555       **11.** Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering  
556       clinical impact with artificial intelligence. BMC Medicine. 2019;17.  
557       <https://doi.org/10.1186/s12916-019-1426-2>
- 558       **12.** Salameh J-P, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred  
559       reporting items for systematic review and meta-analysis of diagnostic test accuracy studies  
560       (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. 2020;370: m2632.  
561       <https://doi.org/10.1136/bmj.m2632>
- 562       **13.** Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2:  
563       A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal  
564       Medicine. 2011;155: 529. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>
- 565

- 566 **14.** Cardoso JR, Pereira LM, Iversen MD, Ramos AL. What is gold standard and what is ground  
567 truth? Dental Press Journal of Orthodontics. 2014;19: 27–30. [https://doi.org/10.1590/2176-](https://doi.org/10.1590/2176-9451.19.5.027-030.ebo)  
568 [9451.19.5.027-030.ebo](https://doi.org/10.1590/2176-9451.19.5.027-030.ebo)
- 569 **15.** Dong X, Du S, Zheng W, Cai C, Liu H, Zou J. Evaluation of an artificial intelligence system for  
570 the detection of diabetic retinopathy in Chinese community healthcare centers. Frontiers in  
571 Medicine. 2022;9. <https://doi.org/10.3389/fmed.2022.883462>
- 572 **16.** Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-  
573 learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA  
574 Ophthalmology. 2019;137: 987. <https://doi.org/10.1001/jamaophthalmol.2019.2004>
- 575 **17.** Hao S, Liu C, Li N, Wu Y, Li D, Gao Q, et al. Clinical evaluation of AI-assisted screening for  
576 diabetic retinopathy in rural areas of Midwest China. Grzybowski A, editor. PLOS ONE.  
577 2022;17: e0275983. <https://doi.org/10.1371/journal.pone.0275983>
- 578 **18.** He J, Cao T, Xu F, Wang S, Tao H, Wu T, et al. Artificial intelligence-based screening for  
579 diabetic retinopathy at community hospital. Eye. 2019;34: 572–576.  
580 <https://doi.org/10.1038/s41433-019-0562-4>
- 581 **19.** Jain A, Krishnan R, Rogye A, Natarajan S. Use of offline artificial intelligence in a smartphone-  
582 based fundus camera for community screening of diabetic retinopathy. Indian Journal of  
583 Ophthalmology. 2021;69: 3150. [https://doi.org/10.4103/ijo.ijo\\_3808\\_20](https://doi.org/10.4103/ijo.ijo_3808_20)
- 584 **20.** Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M-L, Mehrotra A.  
585 Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care.  
586 JAMA Network Open. 2018;1: e182665. <https://doi.org/10.1001/jamanetworkopen.2018.2665>
- 587 **21.** Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, MacIsaac RJ, et al. Feasibility and patient  
588 acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at  
589 endocrinology outpatient services: A pilot study. Scientific Reports. 2018;8.  
590 <https://doi.org/10.1038/s41598-018-22612-2>  
591

- 592       **22.** Li N, Ma M, Lai M, Gu L, Kang M, Wang Z, et al. A stratified analysis of a deep learning  
593           algorithm in the diagnosis of diabetic retinopathy in a real-world study. *Journal of Diabetes*.  
594           2021;14: 111–120. <https://doi.org/10.1111/1753-0407.13241>
- 595       **23.** Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic accuracy of community-  
596           based diabetic retinopathy screening with an offline artificial intelligence system on a  
597           smartphone. *JAMA Ophthalmology*. 2019;137: 1182.  
598           <https://doi.org/10.1001/jamaophthalmol.2019.2923>
- 599       **24.** Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection  
600           in smartphone-based fundus photography using artificial intelligence. *Eye*. 2018;32: 1138–  
601           1144. <https://doi.org/10.1038/s41433-018-0064-9>
- 602       **25.** Scheetz J, Koca D, McGuinness M, Holloway E, Tan Z, Zhu Z, et al. Real-world artificial  
603           intelligence-based opportunistic screening for diabetic retinopathy in endocrinology and  
604           indigenous healthcare settings in Australia. *Scientific Reports*. 2021;11.  
605           <https://doi.org/10.1038/s41598-021-94178-5>
- 606       **26.** Sosale B, Aravind SR, Murthy H, Narayana S, Sharma U, Gowda SGV, et al. Simple, mobile-  
607           based artificial intelligence algorithm in the detection of diabetic retinopathy (SMART) study.  
608           *BMJ Open Diabetes Research & Care*. 2020;8: e000892. [https://doi.org/10.1136/bmjdr-2019-](https://doi.org/10.1136/bmjdr-2019-000892)  
609           [000892](https://doi.org/10.1136/bmjdr-2019-000892)
- 610       **27.** Yang Y, Pan J, Yuan M, Lai K, Xie H, Ma L, et al. Performance of the AIDRScreening system  
611           in detecting diabetic retinopathy in the fundus photographs of Chinese patients: A prospective,  
612           multicenter, clinical study. *Annals of Translational Medicine*. 2022;10: 1088–1088.  
613           <https://doi.org/10.21037/atm-22-350>
- 614       **28.** Zhang Y, Shi J, Peng Y, Zhao Z, Zheng Q, Wang Z, et al. Artificial intelligence-enabled  
615           screening for diabetic retinopathy: A real-world, multicenter and prospective study. *BMJ Open*  
616           *Diabetes Research & Care*. 2020;8: e001596. <https://doi.org/10.1136/bmjdr-2020-001596>  
617

- 618 29. Zhang W, Li D, Wei Q, Ding D, Meng L, Wang Y, et al. The validation of deep learning-based  
619 grading model for diabetic retinopathy. *Frontiers in Medicine*. 2022;9.  
620 <https://doi.org/10.3389/fmed.2022.839088>
- 621 30. Hong Kong Baptist University, Department of Mathematics. Mean variance estimation. In:  
622 [www.math.hkbu.edu.hk](http://www.math.hkbu.edu.hk) [Internet]. 2023 [cited 8 Apr 2023]. Available:  
623 <https://www.math.hkbu.edu.hk/~tongt/papers/median2mean.html>
- 624 31. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test  
625 accuracy evaluations. *Statistics in Medicine*. 2001;20: 2865–2884.  
626 <https://doi.org/10.1002/sim.942>
- 627 32. Hamaddeh N, Van Rompaey C, Metreau E, Eapen SG. New World Bank country classifications  
628 by income level: 2022-2023. In: *World Bank Blogs* [Internet]. World Bank Group; 1 Jul 2022.  
629 Available: [https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-](https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023)  
630 [income-level-2022-2023](https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023)
- 631 33. McMaster University, Evidence Prime. GRADEpro GDT: GRADEpro guideline development  
632 tool [Software]. 2022. Available: [grade.pro](http://grade.pro)
- 633 34. Ting DSW, Cheung CY, Nguyen Q, Sabanayagam C, Lim G, Lim ZW, et al. Deep learning in  
634 estimating prevalence and systemic risk factors for diabetic retinopathy: A multi-ethnic study.  
635 *Nature Partner Journals: Digital Medicine*. 2019;2. <https://doi.org/10.1038/s41746-019-0097-x>
- 636 35. Sheikh A, Bhatti A, Adeyemi O, Raja M, Sheikh I. The utility of smartphone-based artificial  
637 intelligence approaches for diabetic retinopathy: A literature review and meta-analysis. *Journal*  
638 *of Current Ophthalmology*. 2021;33: 219. <https://doi.org/10.4103/2452-2325.329064>
- 639 36. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality  
640 assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-  
641 AI. *Nature Medicine*. 2021;27: 1663–1665. <https://doi.org/10.1038/s41591-021-01517-0>
- 642
- 643 37. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015  
644 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*.  
645 2016;6: e012799. <https://doi.org/10.1136/bmjopen-2016-012799>

646 **38.** Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a  
647 reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the  
648 STARD-AI protocol. *BMJ Open*. 2021;11: e047709. [https://doi.org/10.1136/bmjopen-2020-](https://doi.org/10.1136/bmjopen-2020-047709)  
649 [047709](https://doi.org/10.1136/bmjopen-2020-047709)

650

## 651 **Supporting Information**

652 **S1 Table. Search strategy**

653 **S2 Table. PRISMA-DTA**

654

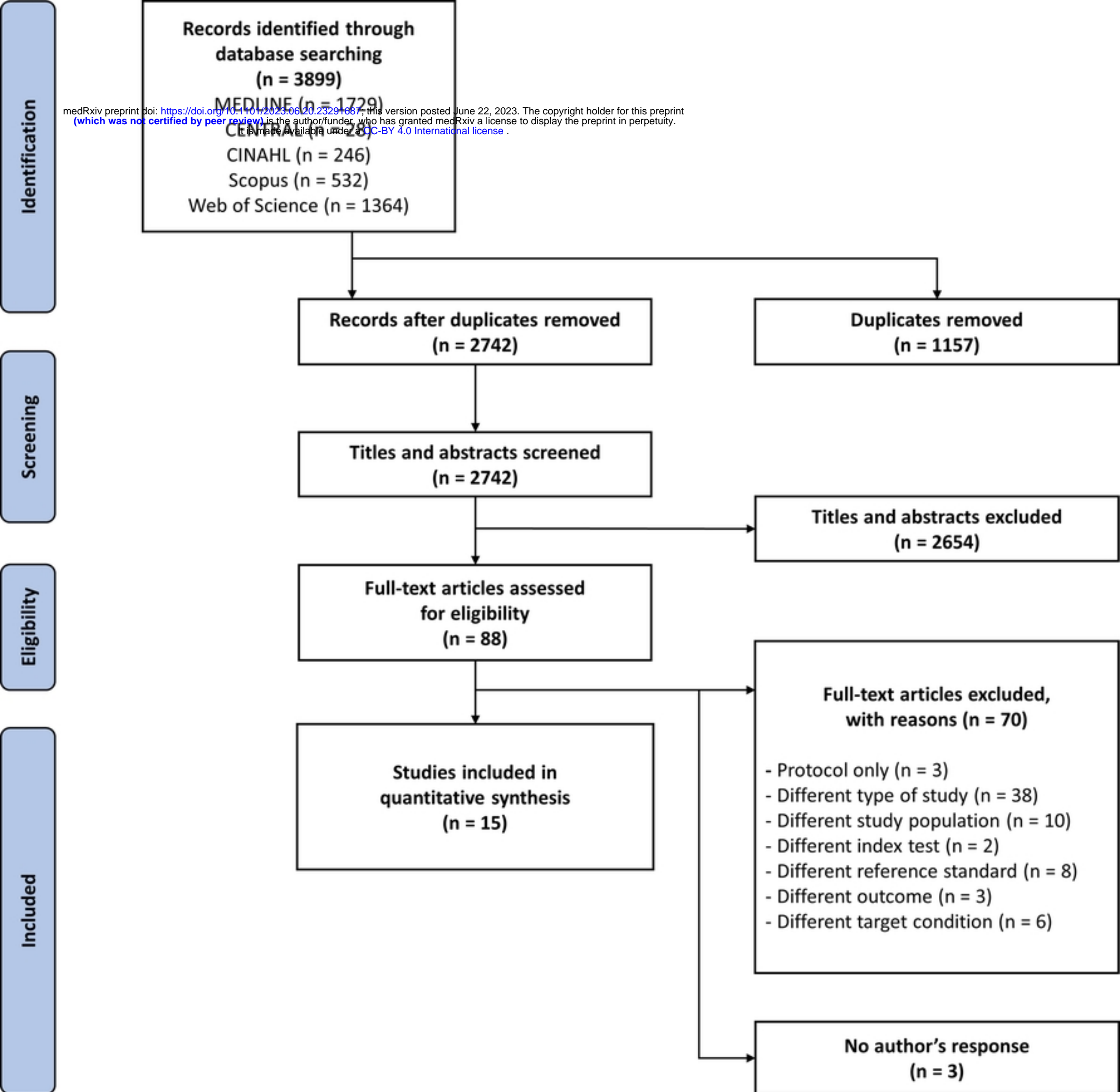


Fig 1



	<u>Risk of Bias</u>				<u>Applicability Concerns</u>		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Dong 2022	?	+	?	+	+	+	+
Gulshan 2019	-	-	+	+	-	+	+
Hao 2022	-	?	+	+	-	+	+
He 2020	?	?	+	+	+	+	+
Jain 2021	?	?	+	-	+	+	+
Kanagasingam 2018	?	-	-	+	+	+	+
Keel 2018	-	-	?	+	+	+	+
Li 2021	?	?	-	+	+	+	+
Natarajan 2019	?	?	-	+	+	+	+
Rajalakshmi 2018	?	?	+	+	+	+	+
Scheetz 2021	?	?	?	+	+	+	+
Sosale 2020	?	-	+	+	+	+	+
Yang 2022	+	-	?	+	+	+	+
Zhang 2020	?	?	?	+	+	+	+
Zhang 2022	?	-	+	+	+	+	+

medRxiv preprint doi: <https://doi.org/10.1101/2023.06.20.23091987>; this version posted June 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).




 High
 Unclear
 Low

Fig 2

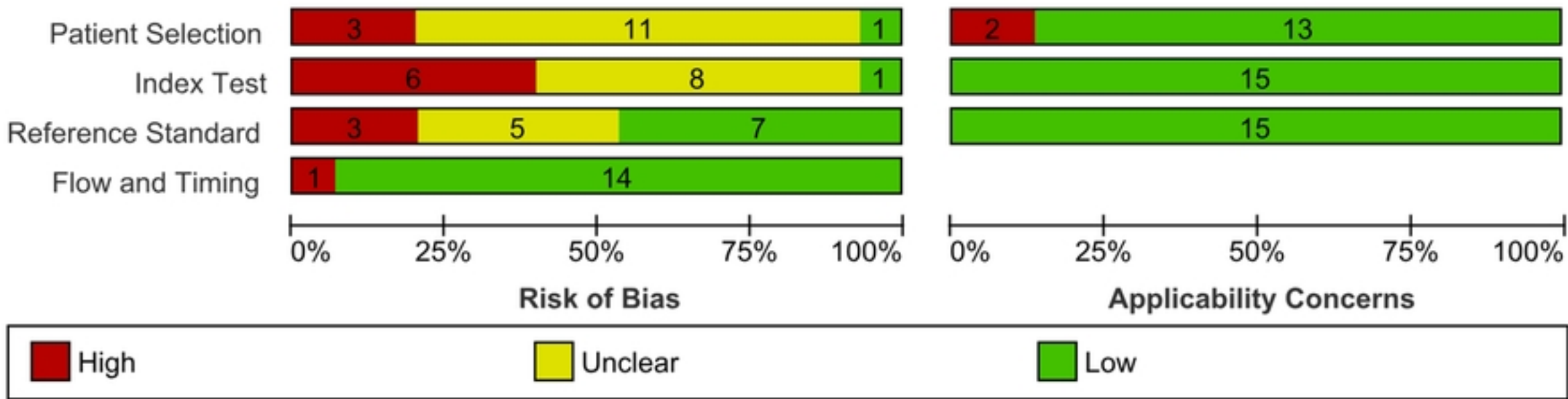


Fig 3

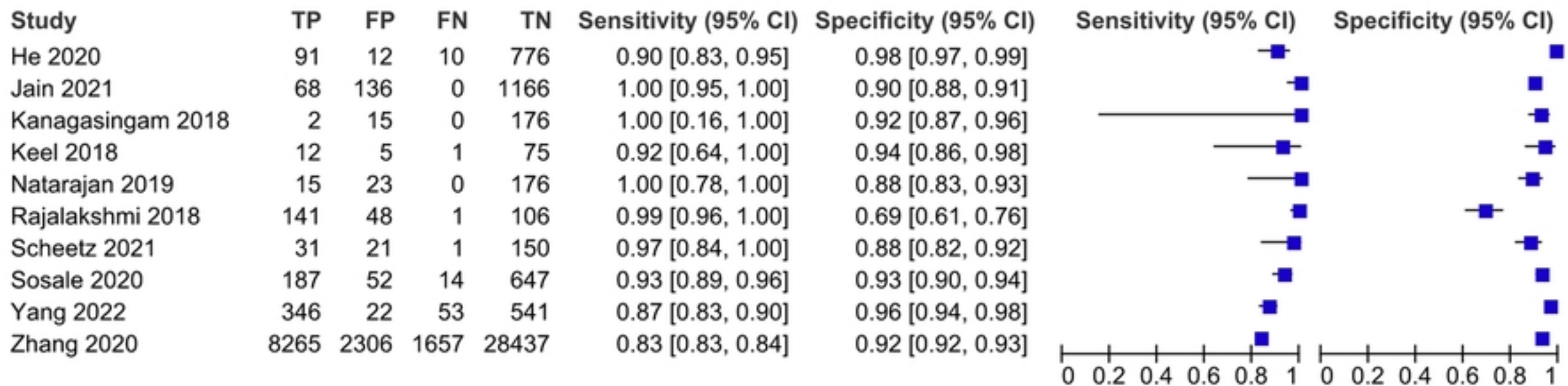


Fig 4

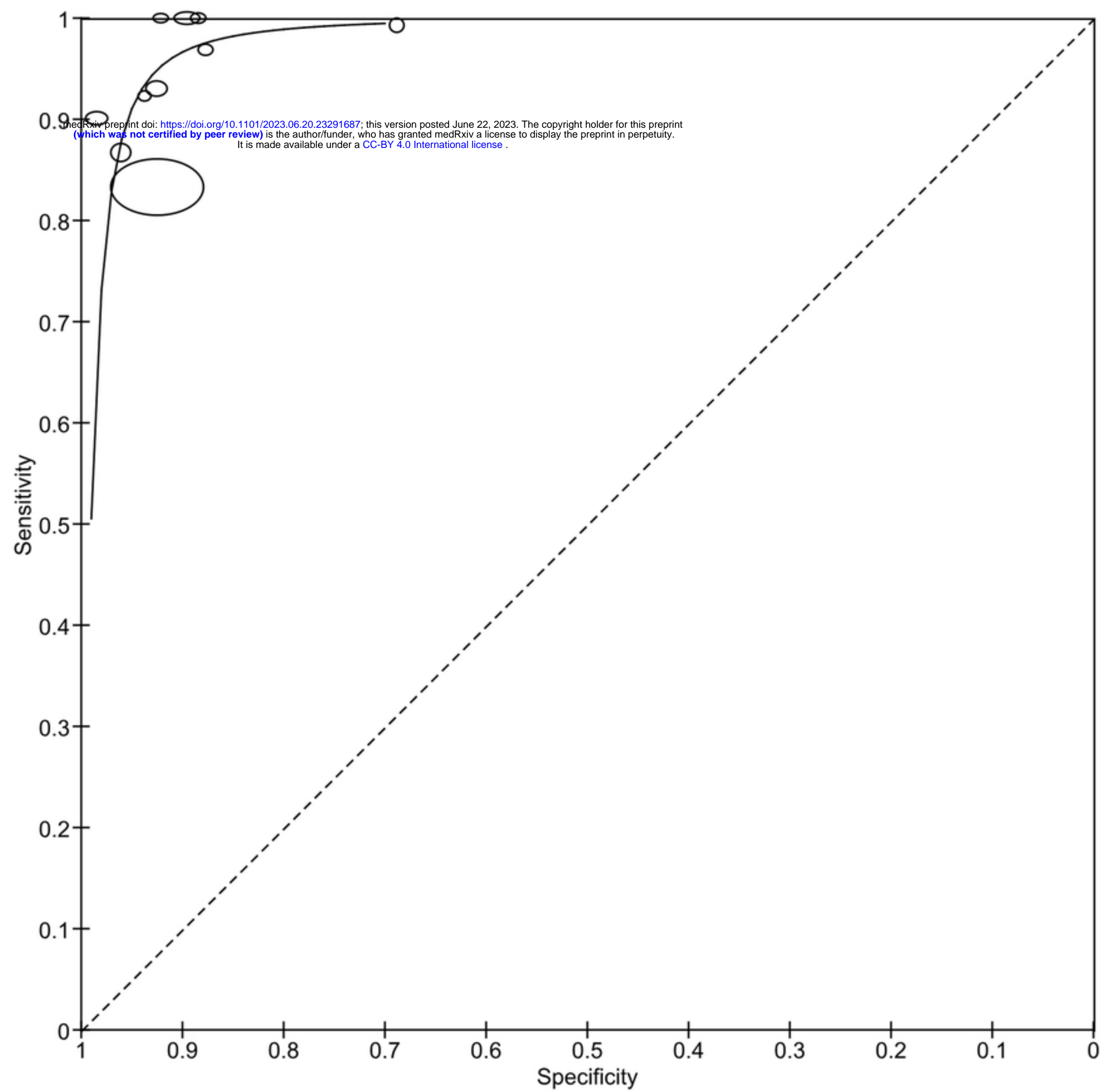


Fig 5

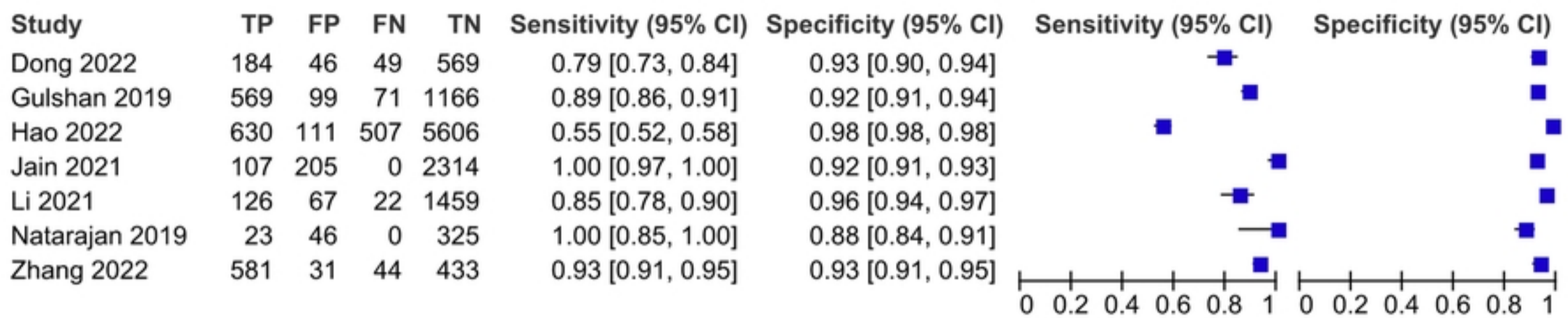


Fig 6

medRxiv preprint doi: <https://doi.org/10.1101/2023.06.20.23291687>; this version posted June 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#).

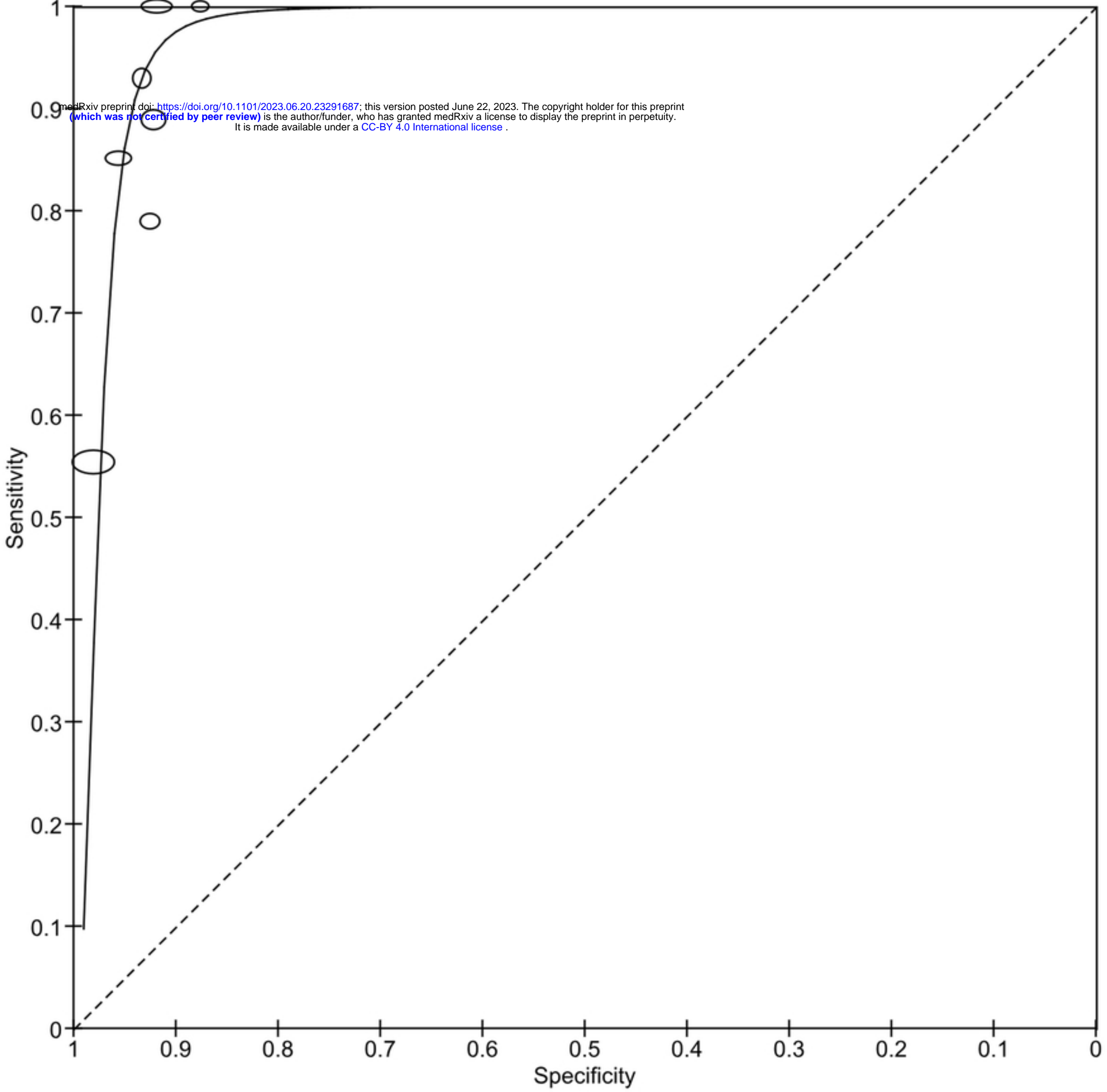
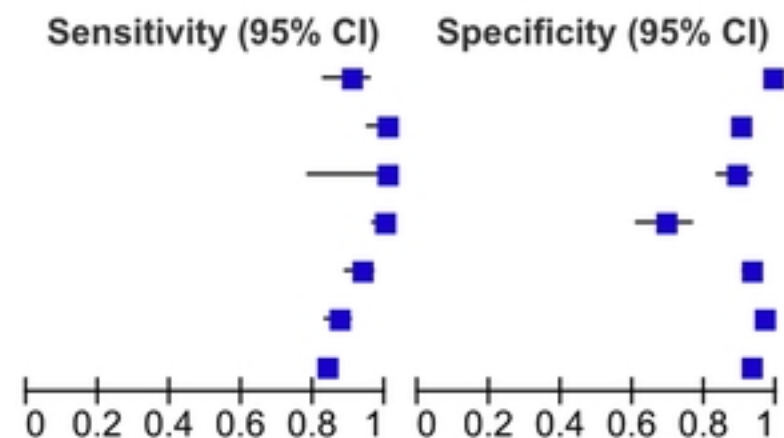


Fig 7

## LMIC

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
He 2020	91	12	10	776	0.90 [0.83, 0.95]	0.98 [0.97, 0.99]
Jain 2021	68	136	0	1166	1.00 [0.95, 1.00]	0.90 [0.88, 0.91]
Natarajan 2019	15	23	0	176	1.00 [0.78, 1.00]	0.88 [0.83, 0.93]
Rajalakshmi 2018	141	48	1	106	0.99 [0.96, 1.00]	0.69 [0.61, 0.76]
Sosale 2020	187	52	14	647	0.93 [0.89, 0.96]	0.93 [0.90, 0.94]
Yang 2022	346	22	53	541	0.87 [0.83, 0.90]	0.96 [0.94, 0.98]
Zhang 2020	8265	2306	1657	28437	0.83 [0.83, 0.84]	0.92 [0.92, 0.93]



## HIC

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Kanagasingam 2018	2	15	0	176	1.00 [0.16, 1.00]	0.92 [0.87, 0.96]
Keel 2018	12	5	1	75	0.92 [0.64, 1.00]	0.94 [0.86, 0.98]
Scheetz 2021	31	21	1	150	0.97 [0.84, 1.00]	0.88 [0.82, 0.92]

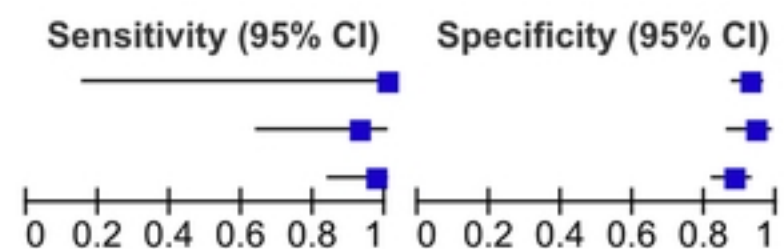
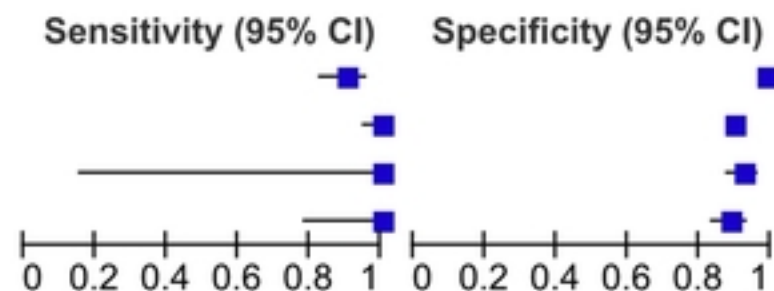


Fig 8

### Primary Level

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
He 2020	91	12	10	776	0.90 [0.83, 0.95]	0.98 [0.97, 0.99]
Jain 2021	68	136	0	1166	1.00 [0.95, 1.00]	0.90 [0.88, 0.91]
Kanagasingam 2018	2	15	0	176	1.00 [0.16, 1.00]	0.92 [0.87, 0.96]
Natarajan 2019	15	23	0	176	1.00 [0.78, 1.00]	0.88 [0.83, 0.93]



### Tertiary Level

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Keel 2018	12	5	1	75	0.92 [0.64, 1.00]	0.94 [0.86, 0.98]
Rajalakshmi 2018	141	48	1	106	0.99 [0.96, 1.00]	0.69 [0.61, 0.76]
Sosale 2020	187	52	14	647	0.93 [0.89, 0.96]	0.93 [0.90, 0.94]
Yang 2022	346	22	53	541	0.87 [0.83, 0.90]	0.96 [0.94, 0.98]

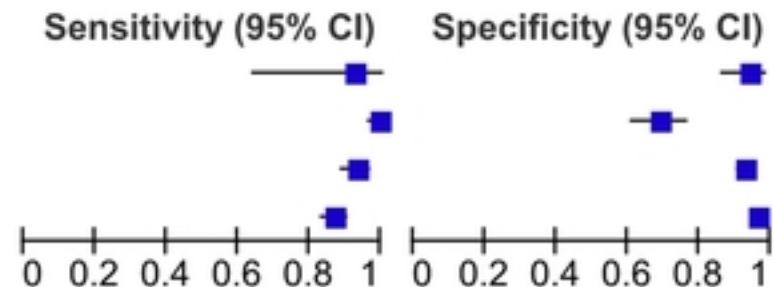


Fig 9



**ICDR**

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
He 2020	91	12	10	776	0.90 [0.83, 0.95]	0.98 [0.97, 0.99]		
Jain 2021	68	136	0	1166	1.00 [0.95, 1.00]	0.90 [0.88, 0.91]		
Kanagasingam 2018	2	15	0	176	1.00 [0.16, 1.00]	0.92 [0.87, 0.96]		
Natarajan 2019	15	23	0	176	1.00 [0.78, 1.00]	0.88 [0.83, 0.93]		
Rajalakshmi 2018	141	48	1	106	0.99 [0.96, 1.00]	0.69 [0.61, 0.76]		
Sosale 2020	187	52	14	647	0.93 [0.89, 0.96]	0.93 [0.90, 0.94]		
Yang 2022	346	22	53	541	0.87 [0.83, 0.90]	0.96 [0.94, 0.98]		
Zhang 2020	8265	2306	1657	28437	0.83 [0.83, 0.84]	0.92 [0.92, 0.93]		

**NHS DES**

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Keel 2018	12	5	1	75	0.92 [0.64, 1.00]	0.94 [0.86, 0.98]		
Scheetz 2021	31	21	1	150	0.97 [0.84, 1.00]	0.88 [0.82, 0.92]		

Fig 10

**DME Included**

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
He 2020	91	12	10	776	0.90 [0.83, 0.95]	0.98 [0.97, 0.99]		
Keel 2018	12	5	1	75	0.92 [0.64, 1.00]	0.94 [0.86, 0.98]		
Natarajan 2019	15	23	0	176	1.00 [0.78, 1.00]	0.88 [0.83, 0.93]		
Rajalakshmi 2018	141	48	1	106	0.99 [0.96, 1.00]	0.69 [0.61, 0.76]		
Scheetz 2021	31	21	1	150	0.97 [0.84, 1.00]	0.88 [0.82, 0.92]		
Sosale 2020	187	52	14	647	0.93 [0.89, 0.96]	0.93 [0.90, 0.94]		

Fig 11

**Human Graders  $\geq 3$** 

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
He 2020	91	12	10	776	0.90 [0.83, 0.95]	0.98 [0.97, 0.99]		
Jain 2021	68	136	0	1166	1.00 [0.95, 1.00]	0.90 [0.88, 0.91]		
Keel 2018	12	5	1	75	0.92 [0.64, 1.00]	0.94 [0.86, 0.98]		
Rajalakshmi 2018	141	48	1	106	0.99 [0.96, 1.00]	0.69 [0.61, 0.76]		
Scheetz 2021	31	21	1	150	0.97 [0.84, 1.00]	0.88 [0.82, 0.92]		
Sosale 2020	187	52	14	647	0.93 [0.89, 0.96]	0.93 [0.90, 0.94]		
Yang 2022	346	22	53	541	0.87 [0.83, 0.90]	0.96 [0.94, 0.98]		
Zhang 2020	8265	2306	1657	28437	0.83 [0.83, 0.84]	0.92 [0.92, 0.93]		

Fig 12