Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: a systematic review and meta-analysis of diagnostic test accuracy

Shirui Wang[1,*], Yuelun Zhang[2,*], Shubin Lei[1,*], Huijuan Zhu[3], Jianqiang Li[4], Qing Wang[5], Jijiang Yang[5], Shi Chen[3] and **Hui Pan[3]**

[1]Eight-year Program of Clinical Medicine, Peking Union Medical College Hospital (PUMCH), Chinese Academe of Medical Sciences & Peking Union Medical College (CAMS & PUMC), Beijing, China, [2]Medical Research Center, PUMCH, CAMS & PUMC, Beijing, China, [3]Department of Endocrinology, Key Laboratory of Endocrinology of National Health Commission, PUMCH, CAMS & PUMC, Beijing, China, [4]School of Software Engineering, Beijing University of Technology, Beijing, China, and [5]Research Institute of Information and Technology, Tsinghua University, Beijing, China

*(S Wang, Y Zhang and S Lei contributed equally to this work)

Correspondence should be addressed to S Chen or H Pan
Email
cs0083@126.com or
panhui20111111@163.com

Abstract

Objective: Automatic diabetic retinopathy screening system based on neural networks has been used to detect diabetic retinopathy (DR). However, there is no quantitative synthesis of performance of these methods. We aimed to estimate the sensitivity and specificity of neural networks in DR grading.

Methods: Medline, Embase, IEEE Xplore, and Cochrane Library were searched up to 23 July 2019. Studies that evaluated performance of neural networks in detection of moderate or worse DR or diabetic macular edema using retinal fundus images with ophthalmologists' judgment as reference standard were included. Two reviewers extracted data independently. Risk of bias of eligible studies was assessed using QUDAS-2 tool.

Results: Twenty-four studies involving 235 235 subjects were included. Quantitative random-effects meta-analysis using the Rutter and Gatsonis hierarchical summary receiver operating characteristics (HSROC) model revealed a pooled sensitivity of 91.9% (95% CI: 89.6% to 94.3%) and specificity of 91.3% (95% CI: 89.0% to 93.5%). Subgroup analyses and meta-regression did not provide any statistically significant findings for the heterogeneous diagnostic accuracy in studies with different image resolutions, sample sizes of training sets, architecture of convolutional neural networks, or diagnostic criteria.

Conclusions: State-of-the-art neural networks could effectively detect clinical significant DR. To further improve diagnostic accuracy of neural networks, researchers might need to develop new algorithms rather than simply enlarge sample sizes of training sets or optimize image quality.

European Journal of Endocrinology (2020) **183**, 41–49

European Journal of Endocrinology

Introduction

Diabetic retinopathy (DR) is a common complication of diabetes affecting one-third of diabetic people (1). In 2010, 0.8 million people were blind and 3.7 million people suffered from moderate to severe visual impairment owing to DR (2). With life expectancy extending around the world, the increased prevalence and duration of diabetes mellitus would lead to an growing population suffered from DR and pose

https://eje.bioscientifica.com
https://doi.org/10.1530/EJE-19-0968

Published by Bioscientifica Ltd.

a huge challenge to the healthcare system and the society (3).

The harmful effect of DR to visual acuity can be effectively minimized by early detection and treatment. A number of developed countries have conducted DR screening programs, in which the severity was assessed and suggestions of examination frequency and treatment were provided accordingly (4). These interventions have been proved to effectively control the incidence of visual impairment (5). However, DR screening requires financial supports, specialized equipment, strict quality control, and highly technical ophthalmologists, which limit the undertaking of screening programs in developing countries. Taking India for instance, only 18–35% of diabetic patients ever received an eye examination (6).

The development of artificial intelligence might help to narrow the gap between the enormous demand for screening and limited healthcare resources (7). Since Hinton's study on deep belief network in 2006 (8), the neural network, which is composed of numerous interconnecting neurons and simulates the thinking of human brain, has evolved quickly and showed high accuracy in image recognition compared with previous algorithms (8). Automatic screening systems based on neural networks have been used to distinguish between no referable DR and referable DR, and the latter would be referred to ophthalmologists for further examination and treatment (9, 10). These automatic systems have good performance and could substantially reduce doctor's workload (11). However, there is no quantitative synthesis of diagnostic accuracy of these methods. Researchers tried different ways, including but not limited to improving image quality, expanding sample sizes, and optimizing algorithms, to raise diagnostic accuracy but got conflicting results (12).

We aimed to conduct a systematic review and meta-analysis to quantitatively analyze the diagnostic accuracy of the neural network in detecting referable DR in patients with diabetes mellitus and investigate the factors that affect diagnostic accuracy.

Methods

Data sources and searches

Medline, Embase, the Cochrane Library (including Cochrane Central Register of Controlled Trials and other databases for technology assessments, economic evaluations, etc.), and IEEE Xplore were used to retrieve the potential eligible studies up to 23 July 2019. Supplementary Table 1 (see section on supplementary materials given at the end of this article) shows the detailed search strategy that combined free texts and MeSH terms relating to the target condition (diabetic retinopathy) and the index tests (neural networks). We also manually checked the reference lists of relevant publications including reviews and commentaries to include eligible studies.

Study selection

Studies were eligible if they met the following pre-specified inclusion criteria: using neural networks to detect referable DR and evaluating the accuracy of it, comparing the index test with ophthalmologists' diagnosis as reference standard, making diagnosis based on retinal images captured by fundus photography without assistance of other medical records, and providing sufficient information for quantitative data synthesis. The definition of referable DR is worse than mild nonproliferative diabetic retinopathy according to the International Clinical Diabetic Retinopathy (ICDR) disease severity scale, or Early Treatment Diabetic Retinopathy Study (ETDRS) level 35 or higher. Diabetic macular edema defined as any hard exudates within 1 disc diameter of the macula according to ETDRS is considered as referable DR as well (13).

Two reviewers independently screened the titles and the abstracts of the citations from the literature search. The full texts of potentially relevant studies were further screened for final inclusion. Disagreements were resolved by discussion between two investigators (S W and S L).

Data extraction and quality assessment

The data were extracted independently by two reviewers including study characteristics (authors and year of publication); characteristics of sample set (data sources, sampling method, age, sex, and resolution of retinal images); characteristics of the index test (algorithms, and number of images used in model training); characteristics of reference standard (diagnostic criteria and number and qualification of ophthalmologists); and accuracy data (number of true positives, true negatives, false positives, and false negatives). If authors in the original studies reported the sensitivity and specificity under multiple thresholds, we extracted the accuracy data under the threshold with the largest Youden's index, defined as the sum of sensitivity and specificity minus 1 (14). When there were only two thresholds reported, the data with higher sensitivity was extracted, since the screening test

required better sensitivity compared to specificity (15, 16). If different diagnostic criteria were compared in the same study (17), we only included the criteria which were most commonly reported in other eligible studies to keep the homogeneity of the meta-analysis. We assessed the risk of bias in patient selection, index test, reference standard, and flow and timing of each study using QUADAS-2 (18). According to the Cochrane Handbook for Diagnostic Tests Review, there was not an universally accepted method to detect publication bias in reviews of diagnostic studies, and the methods used in reviews of randomized controlled trials did not apply to reviews of diagnostic studies, so we did not assess publication of bias in our study (19).

Data synthesis and analysis

Extracted two-by-two data were first graphically showed in the forest plot with the point estimate of sensitivity and specificity and their 95% CIs. Considering the unclear and heterogeneous thresholds for diagnosing DR in different neural network methods, we conducted a quantitative random-effects meta-analysis using the Rutter and Gatsonis hierarchical summary receiver operating characteristics (HSROC) model to combine the SROC curves (19). This method is recommended by the Cochrane Collaboration for the data synthesis of diagnostic accuracy (19). Briefly, the HSROC model assumes that each reported data in the 2-by-2 table represents a point on the ROC from the original research. The ROCs from each original research are sampling estimates of the population-level ROC, and the population-level ROC is the purpose of the data synthesis. Hence, data in the 2-by-2 tables can be used to build a model allowing the fixed-effect and random-effects to estimate the population-level ROC. This ROC is defined as the summary ROC (20), from which all pairs of sensitivity and specificity at different diagnostic thresholds can be retrieved. This model does not require the same threshold as in the original researches, since the variation of the thresholds has been already modeled as the threshold effect (reflected by 'β'). In our review, the combined sensitivity and specificity with corresponding 95% CIs were retrieved from the SROC curves using the largest Youden's index (19). We plotted the combined curve and the optimum diagnostic threshold with corresponding 95% confidence region and 95% prediction region.

Subgroup analyses and meta-regression were used to explore the between-study heterogeneity. We explored the following sources of heterogeneity: image resolution, sample size of training set, diagnostic criteria, and architecture of convolutional neural networks (CNN). All covariates used in subgroup analyses were identified *a priori*. Potential source of heterogeneity was included in the HSROC model as a covariate following the recommendations from the Cochrane Handbook for Diagnostic Tests Review (19). We regarded the factor as a source of heterogeneity if the coefficient of the covariate in the HSROC model was statistically significant. Combined sensitivity and specificity with the corresponding 95% CIs in subgroups were estimated using the model with the covariate.

We performed sensitivity analyses to evaluate the robustness of our main findings by exploring the effect of excluding studies only considering referable DR as target condition without assessing the existence of diabetic macular edema and those not using deep learning method or not reporting the architecture they used. All the data analyses were conducted in RevMan 5.3 and SAS University Edition with two-tailed probability of type I error of 0.05 ($\alpha=0.05$).

Results

Of the 2135 records identified by electronic searches and 19 by hand search, 298 full-text articles were assessed for eligibility and 24 studies in 19 publications met our criteria for inclusion (Supplementary Fig. 1), among which de la Torre *et al*. (21), Gulshan *et al*. (16, 22), Ramachandran *et al*. (23), and Voets *et al*. (15) included two studies using different sample sets in one publication, respectively.

Characteristics of eligible studies

The total number of subjects tested in included studies was 235 235. Fourteen studies used images from public databases (EyePACS, DIARETDB1, Messidor, and Messidor-2) (14, 15, 16, 21, 23, 24, 25, 26, 27, 28, 29) and ten studies obtained their images in local screening programs or hospitals (17, 22, 23, 30, 31, 32, 33, 34, 35). Eleven studies described the demographic characteristics of their study population, of whom the mean age was 54.4 to 63 and the percentage of males was 37.8% to 67.2% (16, 17, 22, 24, 26, 30, 31, 33, 34). Sixteen studies reported resolutions of images ranging from 224×224 to 5184×3456 pixels (14, 15, 16, 21, 23, 24, 25, 27, 28, 31, 32, 33). Apart from two studies not reporting the neural network architecture they used (23) and one study using Bayesian regularisation backpropagation neural network (25), the other 21 studies applied CNN in referable DR detection. Overall, five types of CNN-based models and

systems were reported in 19 studies. Four studies used IDx-DR (17, 24, 30, 34), ten studies used inception architecture (15, 16, 22, 26, 28, 31, 32), two studies used VGGNet (27, 35), one study used net B ensemble network (14), and one study used EyeArt system (29). The number of images used to train the algorithms was reported in 18 studies ranging from 10 000 to 1 665 151 (15, 16, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 35, 36). Fifteen studies defined referable DR as moderate nonproliferative diabetic retinopathy, severe nonproliferative diabetic retinopathy, proliferative diabetic retinopathy, and diabetic macular edema (15, 16, 17, 21, 22, 23, 24, 29, 30, 33, 34), while the other nine studies did not evaluate the existence of macular edema and therefore not include it in the definition of referable DR (14, 21, 25, 26, 27, 28, 31, 32, 35). Supplementary Table 2 shows the detailed characteristics of included studies.

Risk of bias assessment of eligible studies

Supplementary Tables 2 and 3 show the results of the risk of bias assessment of included studies. Regarding to patient selection, risk of bias was low in 11 studies (15, 16, 17, 21, 22, 29, 31, 32, 33, 34), unclear in 11 studies due to the insufficient information describing the sampling method (14, 15, 16, 21, 23, 24, 25, 26, 27, 28), and high in two studies resulting from a enrichment strategy and a case-control design (24, 35). With respect of reference standard, neural networks' screening was based on algorithms without knowledge of the doctors' diagnosis in all studies, though seven studies were judged as high risk of bias owing to a post-specified threshold (14, 15, 22, 23). As for reference standard, risk of bias was low in 14 studies (15, 16, 17, 21, 22, 23, 24, 29, 30, 33, 34) and high in 10 studies for the reason that ophthalmologists made diagnosis referring to neural networks' decisions (2/10) (22, 31) and diabetic macular edema was not included in the reference standard (9/10) (14, 21, 25, 26, 27, 28, 31, 32). Because both artificial intelligence and human experts assessed the same images captured at a same time point, there was no time interval between reference standard and index test. Besides, all images received the same reference standard and were included in the final analysis. Therefore, we considered all studies had a low risk in the domain of flow and timing.

Meta-analysis

Figure 1 shows the paired forest plot for sensitivity and specificity with corresponding 95% CIs for each study. Eligible studies were further combined using the HSROC model, and the SROC curve is shown in Fig. 2 with the 95% confidence region and 95% prediction region. We calculated the following summarized estimates using the HSROC model: sensitivity 91.9% (95% CI: 89.6% to 94.3%), specificity 91.3% (95% CI: 89.0% to 93.5%), positive likelihood ratio 10.5 (95% CI: 7.7 to 13.4), negative likelihood ratio 0.09 (95% CI: 0.06 to 0.12), and diagnostic odds ratio 119.0 (95% CI 60.8 to 177.2).

Sources of heterogeneity

Table 1 shows the detailed results of subgroup analyses exploring the potential source of between-study heterogeneity. We found no association between neural networks' accuracy and image resolution, sample size of training sets, type of CNN models, and diagnostic criteria.

Sensitivity analysis

After excluding nine studies which did not consider diabetic macular edema as target condition

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Abramoff 2016	182	89	6	597	0.97 [0.93, 0.99]	0.87 [0.84, 0.89]
Abramoff 2018	173	65	25	556	0.87 [0.82, 0.92]	0.90 [0.87, 0.92]
Al-Jarrah 2017	40	1	2	46	0.95 [0.84, 0.99]	0.98 [0.89, 1.00]
Bhaskaranand 2019	18917	7193	1803	73797	0.91 [0.91, 0.92]	0.91 [0.91, 0.91]
de la Torre 2019a	1727	1235	179	6859	0.91 [0.89, 0.92]	0.85 [0.84, 0.86]
de la Torre 2019b	465	47	61	627	0.88 [0.85, 0.91]	0.93 [0.91, 0.95]
Gulshan 2016a	666	535	17	7570	0.98 [0.96, 0.99]	0.93 [0.93, 0.94]
Gulshan 2016b	244	91	10	1400	0.96 [0.93, 0.98]	0.94 [0.93, 0.95]
Gulshan 2019a	569	99	71	1166	0.89 [0.86, 0.91]	0.92 [0.91, 0.94]
Gulshan 2019b	1062	125	91	2469	0.92 [0.90, 0.94]	0.95 [0.94, 0.96]
Kanagasingam 2018	2	15	0	176	1.00 [0.16, 1.00]	0.92 [0.87, 0.96]
Krause 2018	204	123	6	1480	0.97 [0.94, 0.99]	0.92 [0.91, 0.94]
Kwasigroch 2018	495	105	41	359	0.92 [0.90, 0.94]	0.77 [0.73, 0.81]
Lam 2018	202	9	16	173	0.93 [0.88, 0.96]	0.95 [0.91, 0.98]
Quellec 2017	1200	378	149	5297	0.89 [0.87, 0.91]	0.93 [0.93, 0.94]
Ramachandran 2018a	11	96	2	376	0.85 [0.55, 0.98]	0.80 [0.76, 0.83]
Ramachandran 2018b	285	90	12	813	0.96 [0.93, 0.98]	0.90 [0.88, 0.92]
Sahlsten 2019	2766	105	321	3926	0.90 [0.88, 0.91]	0.97 [0.97, 0.98]
Ting 2017	1952	5858	205	63881	0.90 [0.89, 0.92]	0.92 [0.91, 0.92]
van der Heijden 2018	50	114	23	711	0.68 [0.57, 0.79]	0.86 [0.84, 0.88]
Verbraak 2019	50	76	13	1154	0.79 [0.67, 0.89]	0.94 [0.92, 0.95]
Voets 2019a	629	1239	65	6857	0.91 [0.88, 0.93]	0.85 [0.84, 0.85]
Voets 2019b	310	394	69	975	0.82 [0.78, 0.86]	0.71 [0.69, 0.74]
Yang 2019	294	6	6	194	0.98 [0.96, 0.99]	0.97 [0.94, 0.99]

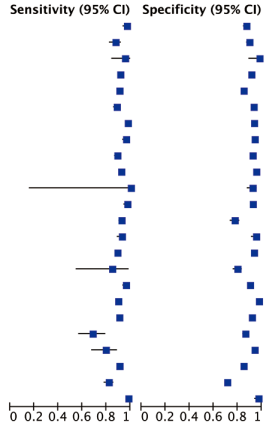

Figure 1
Forest plots of sensitivity and specificity in automatic diagnosis of diabetic retinopathy using neural networks. FN, false negative; FP, false positive; TN, true negative; TP, true positive.

European Journal of Endocrinology

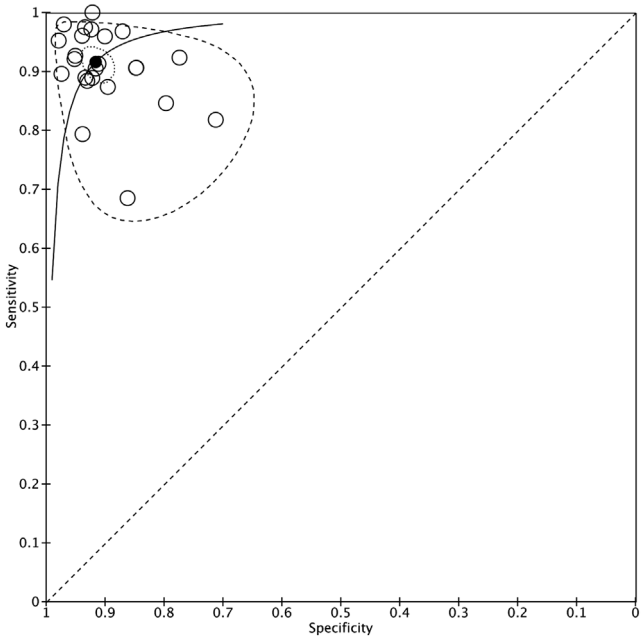

Figure 2
Summary receiver operating characteristics (SROC) curves of eligible studies. Dotted line indicates the 95% confidence region and dashed line indicates the 95% prediction region.

(14, 21, 25, 26, 27, 28, 31, 32), pooled sensitivity was 90.7% (95% CI: 87.2% to 94.3%) and specificity was 90.0% (95% CI: 87.2% to 92.7%). After excluding three studies whose neural networks were unclear or not in-depth (23, 25), pooled sensitivity was 91.7% (95% CI: 89.1% to 94.2%) and specificity was 91.4% (95% CI: 89.1% to 93.8%). These estimates were similar with the main findings in the whole dataset, and hence, we did not find the evidence that the overall combined estimates were influenced by either target condition different from our review questions or non-deep neural networks.

Discussion

To the best of our knowledge, this is the first systematic review and meta-analysis assessing the diagnostic utility of neural networks in DR. Our study shows that the neural network methods can correctly detect 91.9% (95% CI: 89.6% to 94.3%) of the patients with referable DR and exclude 91.3% (95% CI: 89.0% to 93.5%) of patients without referable DR. These results are superior to the pooled sensitivity and specificity demonstrated in previous meta-analyses on computer aided diagnosis of melanoma and breast cancer (37, 38, 39) and implied application prospects of artificial intelligence in DR screening.

We compared the performance of five CNN models in subgroup analyses and found no significant difference among them. The IDx-DR subgroup was attention-worthy as it was the first artificial intelligence system approved by the US Food and Drug Administration (40). Several studies recruiting patients from different background were conducted to assess the performance of it in clinical use, but there have been no studies that summarized those results quantitatively. We provided

Table 1 Subgroup analyses for the accuracy of automatic detection of diabetic retinopathy using neural networks.

Subgroup variables	Number of eligible studies	Sensitivity, % (95 CI)	Specificity, % (95 CI)	*P* for interaction
Image resolution				0.20
<1 million pixels	7	91.8 (88.2 to 95.3)	86.9 (80.8 to 93.0)	
≥1 million pixels	9	92.5 (89.6 to 95.4)	92.5 (89.2 to 95.8)	
Sample size of training set				0.61
<75 000	8	91.6 (87.9 to 95.2)	91.3 (87.0 to 95.7)	
≥75 000	10	93.3 (90.6 to 95.9)	91.4 (87.6 to 95.2)	
CNN model/system				0.57
IDx-DR	5	86.3 (77.5 to 95.2)	89.6 (83.2 to 96.1)	
Inception	10	92.9 (89.7 to 96.0)	92.4 (89.2 to 95.5)	
EyeArt	1	91.3 (80.5 to 100.0)	91.1 (80.1 to 100.0)	
VGGNet	2	96.0 (91.9 to 100.0)	90.4 (81.3 to 99.5)	
net B	1	89.0 (75.5 to 100.0)	93.4 (84.8 to 100.0)	
Diagnostic criteria				0.46
ICDR	19	92.2 (89.5 to 94.8)	91.5 (89.0 to 94.0)	
ETDRS	4	90.6 (83.6 to 97.5)	89.1 (82.4 to 95.7)	

*The image resolution was calculated by multiplying pixel columns and pixel rows. If images of different resolution were used, average resolution was calculated.
CNN, convolutional neural network; ETDRS, Early Treatment Diabetic Retinopathy Study; ICDR, International Clinical Diabetic Retinopathy.

the quantitative evidence of the guaranteed accuracy of IDx-DR using meta-regression. As for other CNN models, our results showed that they might have great potential in clinical application with diagnostic accuracy as good as IDx-DR.

We found no relationship between diagnostic accuracy and image resolution, which contradicted to a previous study on microaneurysms (41). It was generally assumed that high-resolution images could provide more details about lesions and so made it easier to identify (42). Nevertheless, it has been reported that there existed a threshold from which an increase in resolution would not lead to better diagnosis accuracy (42, 43, 44, 45). In addition to image quality, we found no subgroup effect regarding to sample size of training sets. Since it is time-consuming and expensive to accumulate and train algorithms with a large amount of high-resolution labeled data, it might be cost effective if we could find a minimum requirement for image resolution and sample size of training set (46, 47). This finding could be clinically significant from the perspective of rare disease, of which the number of cases is limited. Notwithstanding, it is noteworthy that although there was no significant difference, the subgroup with larger sample size and higher image resolution showed higher sensitivity and specificity. More research is needed before we draw the conclusion regarding to the influence of sample size and image resolution.

The diagnostic accuracy of neural networks might not be influenced by the criteria used in experts' diagnosis, which was reasonable because ICDR was developed on the basis of ETDRS and there is a corresponding relation between these two scales (48). Our results suggested that the agreement between ICDR and ETDRS still exists when it comes to artificial intelligence diagnosis. For the reason that ICDR is easier and more widely used in daily clinical work, it might be suitable to use ICDR as criteria when developing and evaluating computer-assisted screening systems, though ETDRS is regarded as gold standard (48).

Given what was previously mentioned, we found no subgroup effects of the factors commonly supposed to influence the diagnostic accuracy. It suggested that the development of neural networks might have hit a plateau and that it might be difficult to further optimize diagnostic accuracy through existing methods without technology innovation. In November 2019, Xie *et al.* put forward a new self-training model. Unlike common models trained with supervised learning, this new model made use of large quantities of unlabeled images and improved the diagnostic accuracy by one percentage point (49). The emerging new methods provide a probability for diagnostic accuracy improvement.

The diagnostic performance of neural networks on macular edema should be of concern. Although the prevalence of macular edema rises as the severity of DR increases, it actually can be observed in all stages of DR (50). In sensitivity analysis, the high pooled diagnostic accuracy of 15 studies including macular edema in target conditions reflected the capability of neural networks in macular edema detection. Further studies that directly evaluate the diagnostic accuracy of neural networks on macular edema should be done. Moreover, our meta-analysis focused on studies that used fundus image as an examination method, the main shortcoming of which is that it could only produce 2D images, while the structure of retina is actually 3D. From the perspective of techniques, optical coherence tomography (OCT) that could capture the cross-sectional axial of retina through light coherence has an advantage of macular edema visualization and has largely supplanted fundus image in macular edema detection (51, 52). Thus, algorithms should also be trained to interpret OCT images in order to better diagnose macular edema. Furthermore, the artificial intelligence method that could make use of both fundus image and OCT might get better results (52).

We strictly adhered to the guidelines for diagnostic reviews (19). A comprehensive literature search was conducted in not only medical databases but also engineering and technology databases. High-quality and large-scale clinical studies published so far were included in our studies. Sensitivity analysis was conducted to evaluate the robustness of our results. We extracted data that possibly affect the performance of neural networks and image recognition, such as image resolution and characteristics of neural networks, and investigated the heterogeneity resulted from them using meta-regression analysis (41, 53). However, our study had some limitations. First, there were overlapped samples in several studies due to the duplicated data sources used in the original studies. It is difficult to evaluate the influence of the overlapped samples because details of these databases were rarely mentioned in the published paper. Second, the risk of bias was high or unclear in more than half of the studies, especially in the domain of patient selection. It was difficult for us to judge whether consecutive patients were enrolled in some open databases, so the combined sensitivity and specificity might be overestimated due to the possible inconsecutive diseases spectrum. Third, in the subgroup analysis evaluating the effect of CNN models,

only one or two studies contributed data to the VGGNet, EyeArt, and net B subgroups. Since such a small number of studies were included in the analysis, the credibility of our results weakened (54). Fourth, though neural network algorithms have good accuracy in classifying DR severity, they lacked interpretability (55). In other words, the optimum pair of sensitivity and specificity was achieved by optimizing parameters rather than by finding the best cut-off point or threshold. Therefore, since there were no thresholds in traditional sense in these studies, we could not extract and make a comparison of them.

Conclusions

Our review demonstrates the promising performance of neural networks in DR classification using retinal fundus images. Further improvement of diagnostic performance might rely on development of new algorithms rather than only increasing image resolutions or number of images in training sets.

Supplementary materials
This is linked to the online version of the paper at https://doi.org/10.1530/EJE-19-0968.

Declaration of interest
The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this study.

Funding
This study was supported by the Beijing Municipal Natural Science Foundation (7192153), the CAMS Initiative for Innovative Medicine (2016-I2M-1-008), and the National Undergraduates Innovation Training Program of Peking Union Medical Colleges (2019zlgc0607).

Author contribution statement
S C, H P, and J Y designed the study. S W and S L performed the literature search and appraised the articles. S W and Y Z performed the analysis with support from H Z, J L, and Q W. S C, S W, and Y Z wrote the first draft, and H P and J Y finalized the manuscript. All authors reviewed the manuscript and approved the final version of the manuscript.

References

1 Cheung N, Mitchell P & Wong TY. Diabetic retinopathy. *Lancet* 2010 **376** 124–136. (https://doi.org/10.1016/S0140-6736(09)62124-3)
2 Leasher JL, Bourne RR, Flaxman SR, Jonas JB, Keeffe J, Naidoo K, Pesudovs K, Price H, White RA, Wong TY *et al*. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 1990 to 2010. *Diabetes Care* 2016 **39** 1643–1649. (https://doi.org/10.2337/dc15-2171)
3 Wild S, Roglic G, Green A, Sicree R & King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004 **27** 1047–1053. (https://doi.org/10.2337/diacare.27.5.1047)
4 Tozer K, Woodward MA & Newman-Casey PA. Telemedicine and diabetic retinopathy: review of published screening programs. *Journal of Endocrinology and Diabetes* 2015 **2**. (https://doi.org/10.15226/2374-6890/2/4/00131)
5 Antonetti DA, Klein R & Gardner TW. Diabetic retinopathy. *New England Journal of Medicine* 2012 **366** 1227–1239. (https://doi.org/10.1056/NEJMra1005073)
6 Murthy KR, Murthy PR, Kapur A & Owens DR. Mobile diabetes eye care: experience in developing countries. *Diabetes Research and Clinical Practice* 2012 **97** 343–349. (https://doi.org/10.1016/j.diabres.2012.04.025)
7 Mookiah MRK, Acharya UR, Chua CK, Lim CM, Ng EYK & Laude A. Computer-aided diagnosis of diabetic retinopathy: a review. *Computers in Biology and Medicine* 2013 **43** 2136–2155. (https://doi.org/10.1016/j.compbiomed.2013.10.007)
8 Hinton GE & Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006 **313** 504–507. (https://doi.org/10.1126/science.1127647)
9 Li B & Li HK. Automated analysis of diabetic retinopathy images: principles, recent developments, and emerging trends. *Current Diabetes Reports* 2013 **13** 453–459. (https://doi.org/10.1007/s11892-013-0393-9)
10 Ting DS, Cheung GC & Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical and Experimental Ophthalmology* 2016 **44** 260–277. (https://doi.org/10.1111/ceo.12696)
11 A S & S S. Unravelling diabetic retinopathy through image processing, neural networks, and fuzzy logic: a review. *Asian Journal of Pharmaceutical and Clinical Research* 2017 **10** 32–37. (https://doi.org/10.22159/ajpcr.2017.v10i4.17023)
12 Halevy A, Norvig P & Pereira F. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 2009 **24** 8–12. (https://doi.org/10.1109/MIS.2009.36)
13 Photocoagulation for diabetic macular edema: Early Treatment Diabetic Retinopathy Study Report Number 1 Early Treatment Diabetic Retinopathy Study Research Group. *Archives of Ophthalmology* 1985 **103** 1796–1806.
14 Quellec G, Charrière K, Boudi Y, Cochener B & Lamard M. Deep image mining for diabetic retinopathy screening. *Medical Image Analysis* 2017 **39** 178–193. (https://doi.org/10.1016/j.media.2017.04.012)
15 Voets M, Mollersen K & Bongo LA. Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS One* 2019 **14** e0217541. (https://doi.org/10.1371/journal.pone.0217541)
16 Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016 **316** 2402–2410. (https://doi.org/10.1001/jama.2016.17216)
17 van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A & Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmologica* 2018 **96** 63–68. (https://doi.org/10.1111/aos.13613)
18 Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM & QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011 **155** 529–536. (https://doi.org/10.7326/0003-4819-155-8-201110180-00009)

19 Deeks JJ, Bossuyt PM & Gatsonis C. *Handbook for DTA Reviews*, Cochrane Collaboration, London. 2011.
20 Moses LE, Shapiro D & Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993 **12** 1293–1316. (https://doi.org/10.1002/sim.4780121403)
21 de la Torre J, Valls A & Puig D. A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing* 2019 In press. (https://doi.org/10.1016/j.neucom.2018.07.102)
22 Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, Whitehouse K, Coram M, Corrado G, Ramasamy K *et al*. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmology* 2019 137 987-993. (https://doi.org/10.1001/jamaophthalmol.2019.2004)
23 Ramachandran N, Hong SC, Sime MJ & Wilson GA. Diabetic retinopathy screening using deep neural network. *Clinical and Experimental Ophthalmology* 2018 **46** 412–416. (https://doi.org/10.1111/ceo.13056)
24 Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC & Niemeijer M. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology and Visual Science* 2016 **57** 5200–5206. (https://doi.org/10.1167/iovs.16-19964)
25 Al-Jarrah MA & Shatnawi H. Non-proliferative diabetic retinopathy symptoms detection and classification using neural network. *Journal of Medical Engineering and Technology* 2017 **41** 498–505. (https://doi.org/10.1080/03091902.2017.1358772)
26 Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, Peng L & Webster DR. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 2018 **125** 1264–1272. (https://doi.org/10.1016/j.ophtha.2018.01.034)
27 Kwasigroch A, Jarzembinski B & Grochowski M (eds). Deep CNN based decision support system for detection and assessing the stage of diabetic retinopathy. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, 9–12 May 2018, 2018. (https://doi. org/10.1109/IIPHDW.2018.8388337)
28 Lam C, Yi D, Guo M & Lindsey T. Automated detection of diabetic retinopathy using deep learning. *AMIA Joint Summits on Translational Science* 2018 **2017** 147–155.
29 Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda S & Solanki K. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technology and Therapeutics* 2019 **21** 635–643. (https://doi.org/10.1089/dia.2019.0164)
30 Abràmoff MD, Lavin PT, Birch M, Shah N & Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine* 2018 **1** 39. (https://doi.org/10.1038/s41746-018-0040-6)
31 Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML & Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. *JAMA Network Open* 2018 **1** e182665. (https://doi.org/10.1001/jamanetworkopen.2018.2665)
32 Sahlsten J, Jaskari J, Kivinen J, Turunen L, Jaanio E, Hietala K & Kaski K. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific Reports* 2019 **9** 10750. (https://doi.org/10.1038/s41598-019-47181-w)
33 Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017 **318** 2211–2223. (https://doi.org/10.1001/jama.2017.18152)
34 Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO & van der Heijden AA. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care* 2019 **42** 651–656. (https://doi.org/10.2337/dc18-0148)
35 Yang WH, Zheng B, Wu MN, Zhu SJ, Fei FQ, Weng M, Zhang X & Lu PR. An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy and applicability for research. *Diabetes Therapy* 2019 **10** 1811–1822. (https://doi.org/10.1007/s13300-019-0652-0)
36 Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, Tan GSW, Schmetterer L, Keane PA & Wong TY. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology* 2018 **103** 167–175. (https://doi.org/10.1136/bjophthalmol-2018-313173)
37 Rajpara SM, Botello AP, Townend J & Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. *British Journal of Dermatology* 2009 **161** 591–604. (https://doi.org/10.1111/j.1365-2133.2009.09093.x)
38 Dorrius MD, Jansen-van der Weide MC, van Ooijen PM, Pijnappel RM & Oudkerk M. Computer-aided detection in breast MRI: a systematic review and meta-analysis. *European Radiology* 2011 **21** 1600–1608. (https://doi.org/10.1007/s00330-011-2091-9)
39 Noble M, Bruening W, Uhl S & Schoelles K. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. *Archives of Gynecology and Obstetrics* 2009 **279** 881–890. (https://doi.org/10.1007/s00404-008-0841-y)
40 He J, Baxter SL, Xu J, Xu J, Zhou X & Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine* 2019 **25** 30–36. (https://doi.org/10.1038/s41591-018-0307-0)
41 Raman B, Bursell ES, Wilson M, Zamora G, Benche I, Nemeth SC & Soliz P. The effects of spatial resolution on an automated diabetic retinopathy screening system's performance in detecting microaneurysms for diabetic retinopathy. Eds Long R, Antani S, Lee DJ, Nutter B & Zhang M, pp. 128–133, 2004. Proceedings 17th IEEE Symposiumon Computer-Based Medical Systems; 2004 25-25 June 2004.
42 Thapa D, Raahemifar K, Bobier WR & Lakshminarayanan V. Comparison of super-resolution algorithms applied to retinal images. *Journal of Biomedical Optics* 2014 **19** 056002. (https://doi.org/10.1117/1.JBO.19.5.056002)
43 Andreu Y, Lopez-Centelles J, Mollineda RA & Garcia-Sevilla P. IEEE. Analysis of the Effect of Image Resolution on Automatic Face Gender Classification. In *2014 22nd International Conference on Pattern Recognition*, pp. 273–278, 2014. (https://doi. org/10.1109/ICPR.2014.56)
44 Pauli TW, Gangaputra S, Hubbard LD, Thayer DW, Chandler CS, Peng Q, Narkar A, Ferrier NJ & Danis RP. Effect of image compression and resolution on retinal vascular caliber. *Investigative Ophthalmology and Visual Science* 2012 **53** 5117–5123. (https://doi.org/10.1167/iovs.12-9643)
45 Sim DA, Keane PA, Tufail A, Egan CA, Aiello LP & Silva PS. Automated retinal image analysis for diabetic retinopathy in telemedicine. *Current Diabetes Reports* 2015 **15** 14. (https://doi.org/10.1007/s11892-015-0577-6)
46 Panwar N, Huang P, Lee J, Keane PA, Chuan TS, Richhariya A, Teoh S, Lim TH & Agrawal R. Fundus Photography in the 21st Century – a review of recent technological advances and their implications for worldwide healthcare. *Telemedicine Journal and e-Health* 2016 **22** 198–208. (https://doi.org/10.1089/tmj.2015.0068)
47 Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB & Liang J. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Transactions on Medical Imaging* 2016 **35** 1299–1312. (https://doi.org/10.1109/TMI.2016.2535302)
48 Wu L, Fernandez-Loaiza P, Sauma J, Hernandez-Bogantes E & Masis M. Classification of diabetic retinopathy and diabetic macular

edema. *World Journal of Diabetes* 2013 **4** 290–294. (https://doi.org/10.4239/wjd.v4.i6.290)
49 Xie Q, Hov E, Luong M-T & Le QV. Self-training with Noisy Student improves ImageNet classification, 2019. (available at: arXiv:1911.04252)
50 Browning DJ, Fraser CM & Clark S. The relationship of macular thickness to clinically graded diabetic retinopathy severity in eyes without clinically detected diabetic macular edema. *Ophthalmology* 2008 **115** 533–539.e2. (https://doi.org/10.1016/j.ophtha.2007.06.042)
51 Browning DJ, Stewart MW & Lee C. Diabetic macular edema: evidence-based management. *Indian Journal of Ophthalmology* 2018 **66** 1736–1750. (https://doi.org/10.4103/ijo.IJO_1240_18)
52 Hassan B, Hassan T, Li B, Ahmed R & Hassan O. Deep ensemble learning based objective grading of macular edema by extracting clinically significant findings from fused retinal imaging modalities. *Sensors* 2019 **19** E2970. (https://doi.org/10.3390/s19132970)
53 LeCun Y, Bengio Y & Hinton G. Deep learning. *Nature* 2015 **521** 436–444. (https://doi.org/10.1038/nature14539)
54 Richardson M, Garner P & Donegan S. Interpretation of subgroup analyses in systematic reviews: a tutorial. *Clinical Epidemiology and Global Health* 2019 **7** 192–198. (https://doi.org/10.1016/j.cegh.2018.05.005)
55 Niu YH, Gu L, Lu F, Lv FF, Wang ZJ, Sato I, Zhang ZJ, Xiao YY, Dai XZ & Cheng TT. *Pathological Evidence Exploration in Deep Retinal Image Diagnosis*, pp. 1093–1101. Palo Alto: Association Advancement Artificial Intelligence, 2019.

Received 28 November 2019
Revised version received 20 April 2020
Accepted 29 April 2020