Artificial intelligence for diabetic retinopathy in low-income and middle-income countries: a scoping review

Charles R Cleland,1,2 Justus Rwiza,2 Jennifer R Evans,1 Iris Gordon,1 David MacLeod,3 Matthew J Burton,1,4 Covadonga Basaran1

ABSTRACT
Diabetic retinopathy (DR) is a leading cause of blindness globally. There is growing evidence to support the use of artificial intelligence (AI) in diabetic eye care, particularly for screening populations at risk of sight loss from DR in low-income and middle-income countries (LMICs) where resources are most stretched. However, implementation into clinical practice remains limited. We conducted a scoping review to identify what AI tools have been used for DR in LMICs and to report their performance and relevant characteristics. 81 articles were included. The reported sensitivities and specificities were generally high providing evidence to support use in clinical practice. However, the majority of studies focused on sensitivity and specificity only and there was limited information on cost, regulatory approvals and whether the use of AI improved health outcomes. Further research that goes beyond reporting sensitivities and specificities is needed prior to wider implementation.

INTRODUCTION
The application of artificial intelligence (AI) is anticipated to have a considerable impact in many areas of our lives over the coming decades, not least in healthcare. However, implementation of AI into clinical practice remains limited. Ophthalmology is a leading specialty in the development of healthcare AI. In 2018, the first autonomous AI-based medical device to obtain approval from the Food and Drug Administration (FDA) in the USA was the IDx-DR system for the detection and grading of diabetic retinopathy (DR) from retinal photographs. Ophthalmology is a potential exemplar specialty in the application of medical AI, with its use in the context of DR leading the way.

DR is a common complication of diabetes and is a leading cause of blindness globally. DR is often asymptomatic until at an advanced stage, when it is less amenable to treatment; therefore, screening is recommended to prevent sight loss. The early detection and treatment of people with sight-threatening DR substantially reduces the risk of severe visual loss in persons with this stage of disease.

However, many low-income and middle-income countries (LMICs) have no or very limited screening services for DR. As the projected increases in the number of people with diabetes, and consequently DR, will disproportionately affect LMICs, this is a concern. Unless improved screening and treatment services are developed and implemented in LMICs, preventable sight loss from DR will inevitably rise.

A major barrier to the implementation of diabetic eye care services in LMICs is a lack of trained staff. For example, in sub-Saharan Africa (SSA), the region projected to see the largest proportionate increase in the number of people living with DR (up by 143% to 16.3 million people by 2045), there are 2.5 ophthalmologists per million population against a global average of 37.5. Healthcare AI, which task shifts away from human resources for healthcare are most stretched.

The aim of this scoping review is to provide eye care staff, policy makers and researchers with an overview of the literature relating to the use of AI for DR in LMICs to guide clinical trials and the potential implementation of AI tools for DR into clinical pathways.

Research questions
Our research questions are:
1. What AI systems have been used for DR either in, or on data from, populations from LMICs?
2. What are the performance metrics and characteristics of the AI tools used?
Performance metrics include diagnostic accuracy and implementation outcomes (acceptability, fidelity, etc); characteristics include regulatory approvals, technical
specifications, cost information and data management functionality.

The research questions are broad to provide a comprehensive overview of the literature, beyond diagnostic accuracy, in order to guide the use of AI for DR in clinical pathways in LMICs.

METHODS
The study is reported according to the Preferred Reporting Items for Systematic reviews and Meta-analyses extensions for Scoping Reviews guidelines. The protocol was registered on the Open Science Framework repository.

A scoping review was considered the most appropriate methodology for answering the research questions. Our methodological approach was informed by the published guidelines for conducting scoping reviews. The core search concepts for the scoping review were DR, AI and LMICs.

Search strategy and selection criteria
We searched MEDLINE (Ovid), Embase (Ovid), Global Health (Ovid) and the Cochrane Central Register of Controlled Trials on the Cochrane Library on November 29, 2022. A Cochrane Eyes and Vision Information Specialist (IG) developed the search strategies. The searches were constructed using Medical Subject Headings and free-text terms for the following topic areas: “artificial intelligence”, “diabetic retinopathy” and “low- and middle-income countries”. No language limits were applied to the searches. The searches were limited to 2008 onwards. In view of the substantial advances in technology since this date, any publications prior to 2008 are unlikely to be relevant to our objectives. The search strategies are presented in online supplemental appendix 1.

In order to capture studies using imaging data from LMICs, we additionally searched for publications that used 14 publicly available ophthalmic imaging datasets which are from LMICs. A list of these LMIC datasets is available in online supplemental appendix 2. This list was informed by a recent review of all publicly available ophthalmic datasets in *Lancet Digital Health*, which detailed the country of origin of the imaging data.

The inclusion and exclusion criteria were defined before conducting the review but, in keeping with guidelines for scoping reviews, articles were selected during the title and abstract screening if (1) they referred to the use of AI in the context of DR and (2) were conducted in, or using data from people living in an LMIC. All primary research studies were included.

Reviews were excluded but their reference lists were searched for any primary articles that were not included from our original search. Gray literature and conference abstracts were excluded as they do not provide sufficient evidence to inform clinical trials or practice.

AI was defined as any technology, computer software or algorithm that makes an autonomous decision in a manner that mimics human cognition. DR is a complication of diabetes and we included articles that discussed DR or diabetic macular edema (DME). LMICs were defined according to the World Bank definition for 2021.

The differences between our protocol and the review included using the 2021 World Bank definition of LMIC as opposed to 2019 definition stated in our protocol and the addition of a google search to identify additional relevant information about the identified AI systems.

Selection of studies
All identified records were imported into Covidence (Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia, available at www.covidence.org) for screening. Two authors (CRC and CB) independently reviewed each title and abstract and excluded those not meeting the inclusion criteria. Disagreements were resolved by discussion and consensus. The full texts were then again reviewed independently by two reviewers (CRC and CB) to determine which articles should be included in the data extraction phase, with all disagreements resolved by discussion and consensus.

Data charting process
A data extraction form was developed in Covidence based on the scoping review questions and was piloted by two reviewers (CRC and CB). The form was refined based on discussion and finalized prior to extraction. Data extraction was then carried out for each publication independently by two reviewers (CRC and JR). After extraction, all differences were resolved by discussion and consensus.

Data items
Characteristics of publication:
- Title, year of publication, journal
- Affiliation of first author
- Sources of funding
- Stated conflicts of interests for any authors

Characteristics of the AI tool:
- Stated name of AI tool
- Function/Intended use of AI tool
- Would the published study be considered an external validation
- Regulatory approvals
- Purchase cost of AI software
- DR classification system used

External validation was defined as the testing of AI on a new set of data entirely separate from the training dataset. This is a crucial step in the development of AI as it demonstrates that an AI model can work in patients and populations external to the development population. A Google search of the identified AI systems was performed to extract information on the regulatory approvals and purchase cost of the AI not accessible from the publications.

Characteristics of data used to assess performance of AI:
Emerging technologies, pharmacology and therapeutics

- Type of imaging data used (retinal photographs, optical coherence tomography (OCT))
- Country of origin of data
- Data collected retrospectively or prospectively
- Details of reference standard and arbitration process
- Reported performance of AI tool:
  - Sensitivity, specificity, area under the curve (AUC)
- Economic evaluation outcomes
- Implementation research outcomes (fidelity, acceptability, adoption, sustainability)
- Any other reported outcome data not already captured

Synthesis of the results

We conducted a descriptive analysis of the study characteristics, study methods and of the AI tools. Study characteristics included the location of the study (defined as where the imaging data was from), year of publication, funding and conflicts of interest and first author affiliation. The study methods captured whether the data used to train/test the AI were collected prospectively or retrospectively, how many images or participants were included in the dataset, who provided the reference standard and details of the arbitration process. For the AI tool, we captured what task the AI was designed to perform, the performance of the AI for its given task, the name and/or developer if stated, regulatory or cost data and any implementation research outcomes.

The location of all included studies was displayed visually on a map. The studies were then coded into those that were, and were not, considered external validations; key data of studies that were considered external validations were summarized and tabulated. Externally validated studies were then coded into those that had a named AI or a stated developer and those that did not. For those with a named AI and/or stated developer, the sensitivity, specificity and AUC of the AI tool in detecting referable DR was tabulated along with the reference standard, arbitration process and any other key outcome measures.

A consultation stage, which is considered optional in scoping reviews, was not undertaken.17

RESULTS

Figure 1 shows a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow chart outlining the selection process for the included articles.

The searches last run on November 29, 2022 retrieved a total of 521 records. After 121 duplicate records were removed, 400 unique records were screened by title and abstract. A total of 279 records were excluded at the title and abstract stage and 121 records went forward to full-text review. Two reports could not be sourced, therefore a total of 119 reports of studies were assessed for potential inclusion in the review. After reading the full texts, 74 studies were included and 45 studies were excluded.

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram. LMIC, low-income and middle-income country.
The reference lists of 62 reviews were assessed and an additional 7 studies not identified in our primary searches were included. Therefore, a total of 81 studies met our inclusion criteria and were included in the analysis.

**Characteristics of the publications**

The majority of the identified studies were undertaken in three countries: India, China and Thailand (figure 2) with 65% (n=53) of publications in 2020, 2021 and 2022. The majority of first author affiliations were from institutions in LMICs (n=62; 77%); however, of the studies conducted in SSA (n=7; 9%), only one first author was affiliated with an institution in an African country.18

The primary aim of most studies (n=72; 89%) was to assess performance of the AI for a specific task. These studies either focused on a description of AI tool development and performance in the training dataset (n=29; 36%) or a description of performance in an external dataset (n=43; 53%) that was entirely separate from the training data; therefore, meeting our definition of external validation.

**Characteristics of the AI tools**

The function of the majority of the AI tools identified (n=49; 60%) was to automate retinal photograph interpretation and produce a DR grade. Of the remaining studies, five focused on using AI for automating the interpretation of OCT imaging for DME19–23 and one study evaluated the performance of AI in fundus fluorescein angiography interpretation.24 Two studies considered the performance of an AI tool for grading DR and multiple other retinal conditions25 26 and a further two studies considered the performance of two different AI models for grading DR, possible glaucoma and age-related macular degeneration/referable macular diseases.27 28

One study used an AI tool to identify macular edema from two-dimensional retinal photographs29; one study used an AI tool to identify, in persons with diabetes, fundus images without DR (ie, normal fundi)30; one study used heat maps to aid with the ‘black box’ phenomenon in an attempt to understand why an AI model might produce false positives in the context of DR grading31 and one study used AI to inform the photographer whether images taken with a handheld smartphone without mydriasis were gradable for DR and to assess if this could reduce the number of ungradable images captured.32

Seven studies focused on identifying specific features on a retinal image, such as the presence of hard exudates or vessel bifurcation, which are informative when grading an image for DR.33–39

Two studies reported the use of an AI tool to predict the likelihood of DR progression40 41; one study assessed whether AI-assisted image grading can improve human grading42; two studies assessed the impact of using an AI model on patient flow within DR screening services43 44 and one study used an AI model to collect DR prevalence data.45

Two studies had an implementation research focus46 47 and two studies evaluated the cost-effectiveness of using AI for DR screening.48 49

**Externally validated studies**

A total of 43 studies (52%) met our definition of external validation. In summary, the majority (n=33; 77%) reported the use of AI for automating the interpretation...
of retinal images and producing a DR grade and 25 (58%) used prospectively collected data.

Thirty-one (72%) of the externally validated studies had a named AI tool and/or details of the company that developed the tool; therefore, 12 studies did not state the AI name or details of the developer. Of those studies with a named AI and/or developer details, 16 had direct declared conflicts of interests relating to either the AI software or the company that developed the AI and 9 of those studies were funded by the company that developed the AI used in the respective study.

Only four studies stated that the AI tool assessed was commercially available,50–52 however, the company referred to in two of these publications (Visulytix) has ceased trading.

**AI model performance**

As noted above, the majority of studies (n=50; 61%) reported the use of an AI tool for automated retinal image interpretation in order to produce a DR grade. Table 1 details the performance of the externally validated AI models which also had a stated name and/or developer for the detection of referable DR. The sensitivities and specificities for the detection of referable DR in these studies ranged from 83.3%–100% to 68.8%–98%, respectively (figure 3). One study reported that AI performance was not significantly affected by gender.53

As Visulytix, the company responsible for developing the Pegasus AI system, has ceased trading, those two studies were not included in table 1. One study relating to the DAPHNE AI system was also not included in table 1 as it identified fundus images for the presence of microaneurysms only.

The majority of studies used the International Clinical Diabetic Retinopathy Severity Scale,44 classification system with a threshold of moderate NPDR or worse defining referable DR. A clear description of who determined the reference standard DR grade along with the arbitration process was provided by the majority of studies, although the arbitration methodology differed between studies (see table 1 and online supplemental table 1).

The majority of studies excluded images deemed ungradable by the human graders from the analyses. However, five studies did compare images deemed ungradable by human graders with the AI model gradings. These studies all showed that AI tools considered a higher number of images as ungradable when compared with the human graders.52,55–58

Two of the five studies that tested the performance of an AI tool for detecting DME from OCT imaging met our definition of external validation. The Pegasus-OCT system detected age-related macular degeneration (AMD) and DME with a minimum area under the receiver operating characteristic of 99% and 98%, respectively and Tang et al reported an AUC of >0.906 in all the external test sets for detecting the presence of DME.50

**Other outcome measures**

Three studies reported that the AI tool was a registered medical device in China,45 58 59 and one study reported that the AI tool was registered as a class IIa medical device.50 No study stated the cost of the AI tool.

We identified two economic evaluation studies undertaken in China and Brazil.48 49 AI was found to be more cost-effective than the standard of care in the study undertaken in China but not in Brazil.

Our Google search of named AI systems or those with a stated developer revealed commercial websites for the Medios AI (Remido Innovative Solutions),60 SELENA+ (EyRIS),61 EyeArt (Eyenuk),62 RAIDS (SightAI Technology)63 and EyeWisdom (Visionary Intelligence (Vistel))64 systems. Cybersight is provided as a free (non-commercial) software by Orbis International.65 No website provided any cost information other than Cyber Sight/Orbis, which stated their software is free to use in LMICs.

EyeArt’s website stated their AI system has US FDA clearance, CE marking as a class IIa medical device in the European Union and a Health Canada license and the SELENA+/EyRIS website stated that their software has both a Health Science Authority certification from Singapore and is CE marked. There were no details of any regulatory approvals on the Medios AI, RAIDS, EyeWisdom or Cybersight/Orbis website.

Three studies reported implementation research outcomes. A study undertaken in Brazil discussed the feasibility of using AI for DR screening and mainly highlighted the need to raise awareness of diabetic eye disease within the population.47 The second study was undertaken in Thailand when Google’s AI system was implemented in an active clinical pathway.46 The paper reported challenges with fidelity, particularly highlighting poor internet connectivity and suboptimal lighting affecting retinal image capture as issues, as well as acceptability concerns from nursing staff involved in DR screening. A third study in Rwanda demonstrated that for persons screened for DR, AI with a point-of-care referral decision significantly increased the proportion of persons referred from screening who attended the referral eye clinics.18

One study used deep learning to predict the 2-year progression from no DR on retinal imaging to signs of DR. The reported AUC was 0.70 (95% CI 0.67 to 0.74) using the deep learning model alone and this increased marginally to 0.71 (95% CI 0.68 to 0.75) when additional clinical risk factors, notably hemoglobin A1c, were added to the model.40

Two studies assessed the performance of an AI tool for the detection of multiple retinal pathologies, including DR.25 27 The SELENA+ AI system can detect DR, possible glaucoma and AMD. Ting et al reported an AUC for the SELENA+ AI system of 0.942 (95% CI 0.929 to 0.954), sensitivity of 96.4% (95% CI 81.7% to 99.9%) and specificity of 87.2% (95% CI 86.8% to 87.5%) for possible glaucoma and an AUC of 0.931 (95% CI 0.928 to 0.935),
### Table 1: Reported sensitivity, specificity and AUC for the detection of referable DR for externally validated studies that state either the name of the AI tool or the developer

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>AI system/developer</th>
<th>Study type</th>
<th>Number of participants or images</th>
<th>Who provided the reference standard</th>
<th>DR classification system used</th>
<th>Sensitivity (95% CI) for referable DR</th>
<th>Specificity (95% CI) for referable DR</th>
<th>AUC (95% CI) for referable DR</th>
<th>Referable DR threshold</th>
<th>Other reported outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sosale et al.</td>
<td>India</td>
<td>Medios AI</td>
<td>Prospective</td>
<td>922 participants</td>
<td>Five retinal specialists</td>
<td>ICDRS</td>
<td>93% (91.3 to 94.7)</td>
<td>92.5% (90.8 to 94.2)</td>
<td>0.88</td>
<td>Moderate NPDR or worse and/or presence of DME</td>
<td>None</td>
</tr>
<tr>
<td>Bahl and Rao</td>
<td>India</td>
<td>Medios AI</td>
<td>Prospective</td>
<td>250 participants</td>
<td>None</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>197gradable images obtained from 250 participants</td>
</tr>
<tr>
<td>Natarajan et al.</td>
<td>India</td>
<td>Medios AI</td>
<td>Prospective</td>
<td>231 participants</td>
<td>Two ophthalmologists</td>
<td>ICDRS</td>
<td>100%</td>
<td>88.4%</td>
<td>NS</td>
<td>Moderate NPDR or worse and/or presence of DME</td>
<td>None</td>
</tr>
<tr>
<td>Jain et al.</td>
<td>India</td>
<td>Medios AI</td>
<td>Prospective</td>
<td>1378 participants</td>
<td>Two vitreo-retinal surgeons</td>
<td>ICDRS</td>
<td>100% (94.7 to 100)</td>
<td>89.55% (87.76 to 91.16)</td>
<td>NS</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Ting et al.</td>
<td>China, Australia, Singapore, USA</td>
<td>SELENA+</td>
<td>Retrospective</td>
<td>18913 participants</td>
<td>Ophthalmologist and retina specialist</td>
<td>ICDRS</td>
<td>92.25% (90.10 to 94.12)</td>
<td>89.04% (87.85 to 90.28)</td>
<td>0.963 (0.956 to 0.969)</td>
<td>Moderate NPDR or worse (macular signs not specified)</td>
<td>DLS system was quicker at performing grading than humans</td>
</tr>
<tr>
<td>Bellemo et al.</td>
<td>Zambia</td>
<td>SELENA+</td>
<td>Retrospective</td>
<td>1574 participants</td>
<td>Two trained graders</td>
<td>ICDRS</td>
<td>92.86% (91.76 to 93.96)</td>
<td>90.82% (89.75 to 91.89)</td>
<td>0.973 (0.969 to 0.978)</td>
<td>Moderate NPDR or worse, DME or an ungradable image</td>
<td>None</td>
</tr>
<tr>
<td>Ting et al.</td>
<td>China, Mexico, Singapore</td>
<td>SELENA+</td>
<td>Retrospective</td>
<td>14880 participants (in validation set)</td>
<td>Varies between datasets</td>
<td>ICDRS</td>
<td>90.5% (87.3 to 93.0)</td>
<td>91.6% (91.0 to 92.2)</td>
<td>0.936 (0.925 to 0.943)</td>
<td>Moderate NPDR or worse, DME or an ungradable image</td>
<td>DLS performance for detecting possible glaucoma and AMD also reported</td>
</tr>
<tr>
<td>Sayres et al.</td>
<td>India and USA</td>
<td>Google</td>
<td>Retrospective</td>
<td>1612 participants</td>
<td>Three retinal specialists</td>
<td>ICDRS</td>
<td>91.5%</td>
<td>94.7%</td>
<td>NS</td>
<td>Moderate or worse DR (macular signs not specified)</td>
<td>Human graders assisted by AI improved grading accuracy</td>
</tr>
<tr>
<td>Raumviboonsuk et al.</td>
<td>Thailand</td>
<td>Google</td>
<td>Retrospective</td>
<td>7517 participants</td>
<td>Two retinal specialists</td>
<td>ICDRS</td>
<td>96.8% (95.3 to 98.3)</td>
<td>95.6% (95.4 to 98.7)</td>
<td>98.4% (96.7 to 99.5)</td>
<td>Publication did not specify RDR, these figures are taken for the detection of moderate NPDR or worse. Referable DME data were also reported</td>
<td>Algorithm was statistically significantly better than local Thai graders</td>
</tr>
<tr>
<td>Gulshan et al.</td>
<td>India</td>
<td>Google</td>
<td>Prospective</td>
<td>3049 participants</td>
<td>A trained grader and a retinal specialist</td>
<td>ICDRS</td>
<td>Aravind: 88.9% (95.8 to 91.1) Sankara: 92.1% (90.1 to 93.8)*</td>
<td>Aravind: 92.2% (90.3 to 93.8) Sankara: 95.2% (94.2 to 96.1).*</td>
<td>Aravind: 0.963. Sankara: 0.980</td>
<td>Moderate NPDR or worse and/or presence of DME</td>
<td>None</td>
</tr>
<tr>
<td>Gulshan et al.</td>
<td>India and USA</td>
<td>Google</td>
<td>Retrospective</td>
<td>Does not specify how many images from India used to train model!</td>
<td>Seven and eight ophthalmologists graded the images from the Messidor-2 and EyePACS-1 datasets, respectively</td>
<td>ICDRS</td>
<td>EyePACS-1: 97.5% (95.8 to 98.7) Messidor-2: 94.1% (92.4 to 95.8)†</td>
<td>EyePACS-1: 93.4% (92.8 to 94.0) Messidor-2: 93.9% (92.4 to 95.3)†</td>
<td>0.991 (0.988 to 0.995)†</td>
<td>Moderate NPDR or worse and/or presence of DME</td>
<td>Algorithm performance tested at different cut-offs (high sensitivity, high specificity, etc)</td>
</tr>
<tr>
<td>Boris et al.</td>
<td>Thailand and USA</td>
<td>Google</td>
<td>Retrospective</td>
<td>6791 participants</td>
<td>One ophthalmologist graded each image from a pool of 11</td>
<td>ICDRS</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>AUC for predicting risk of developing DR at 2 years from normal fundus image at baseline 0.70 (0.67 to 0.74)</td>
</tr>
</tbody>
</table>

*Aravind: 88.9% (95.8 to 91.1) Sankara: 92.1% (90.1 to 93.8)
†EyePACS-1: 97.5% (95.8 to 98.7) Messidor-2: 94.1% (92.4 to 95.8)
‡EyePACS-1: 93.4% (92.8 to 94.0) Messidor-2: 93.9% (92.4 to 95.3)
<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>AI system name/developer</th>
<th>Study type</th>
<th>Number of participants or images</th>
<th>Who provided the reference standard</th>
<th>DR classification system used</th>
<th>Sensitivity (95% CI) for referable DR</th>
<th>Specificity (95% CI) for referable DR</th>
<th>AUC (95% CI) for referable DR</th>
<th>Referable DR threshold</th>
<th>Other reported outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruamviboonsuk et al(^{55})</td>
<td>Thailand</td>
<td>Google</td>
<td>Prospective</td>
<td>7651 participants</td>
<td>Three retinal specialists</td>
<td>ICDRS</td>
<td>95.1% (91.8 to 97.9)</td>
<td>89.3% (85.8 to 92.3)</td>
<td>NS</td>
<td>Moderate NPDR or worse, DME or an ungradable image</td>
<td>Qualitative interviews revealed that nurses positively viewed the system but concerns were expressed about images deemed ungradable by the DLS and the additional time required for image upload</td>
</tr>
<tr>
<td>Al-Turk et al(^{56})</td>
<td>Kenya, China, Saudi Arabia</td>
<td>DAPHNE</td>
<td>Retrospective</td>
<td>Kenya: 24700 images, China: 15000 images</td>
<td>China: NS Kenya: Moorfields Eye Hospital retinal image reading center</td>
<td>Kenya: ICDRS China: UK NSC guidelines</td>
<td>Kenya: 94.28% (93.1 to 95.2) China: 95.51% (93.1 to 97.5)</td>
<td>Kenya: 92.12% (88.27 to 93.33) China: 91.11% (89.11 to 92.63)</td>
<td>NS</td>
<td>UK NSC guideline and ICDRS (moderate or worse DR) classifications used to determine referral threshold including presence of DME</td>
<td>None</td>
</tr>
<tr>
<td>Al-Turk et al(^{57})</td>
<td>Kenya, China, Saudi Arabia</td>
<td>DAPHNE</td>
<td>Retrospective</td>
<td>49726 images</td>
<td>NS</td>
<td>Kenya: ICDRS China: UK NSC guidelines</td>
<td>Kenya: 92.03% (91.2 to 92.9) China: 92.9% (91.4 to 94.2)</td>
<td>Kenya: 93.0% (92.7 to 93.4) China: 94.4% (94.1 to 97.9)</td>
<td>NS</td>
<td>Kenya: moderate NPDR or worse using ICDRS criteria and/or presence of DME, China: UK NSC cut-off used (includes presence of DME)</td>
<td>Ability of DAPHNE to detect non-DR pathology was assessed; sensitivity and specificity data not reported due to small number of cases</td>
</tr>
<tr>
<td>Rajalakshmi et al(^{58})</td>
<td>India</td>
<td>EyeArt software (V.2.1.0)</td>
<td>Prospective</td>
<td>301 participants</td>
<td>Two ophthalmologists</td>
<td>ICDRS</td>
<td>99.3% (96.1 to 99.9)</td>
<td>68.8% (61.5 to 76.2)</td>
<td>NS</td>
<td>Moderate NPDR or worse and/or presence of DME</td>
<td>None</td>
</tr>
<tr>
<td>Zhang et al(^{59})</td>
<td>China</td>
<td>VolexCloud</td>
<td>Prospective</td>
<td>15605 participants</td>
<td>Two specialist graders</td>
<td>ICDRS</td>
<td>83.3% (81.9 to 84.6)</td>
<td>92.5% (92.1 to 92.9)</td>
<td>NS</td>
<td>Moderate NPDR or worse (DME assessed separately)</td>
<td>Prevalence data in Chinese population reported</td>
</tr>
<tr>
<td>Li et al(^{60})</td>
<td>China</td>
<td>VolexCloud</td>
<td>Prospective</td>
<td>11476 participants</td>
<td>One certified retinal specialist</td>
<td>ICDRS</td>
<td>85.1% (83.4 to 86.6)</td>
<td>90.6% (94.6 to 96.6)</td>
<td>0.942 (0.920 to 0.964)</td>
<td>Moderate NPDR or worse (macular signs not specified)</td>
<td>389 images were ungradable using DLA but were gradable by humans</td>
</tr>
<tr>
<td>Lin et al(^{61})</td>
<td>China</td>
<td>Comprehensive AI Retinal Expert</td>
<td>Prospective</td>
<td>18163 images in external validation dataset</td>
<td>Three ophthalmologists</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>$1.0.960 (0.993 to 0.966) 2.0.999 (0.998 to 1.000) 3.0.918 (0.887 to 0.944)$</td>
<td>Moderate NPDR or worse (macular signs not specified)</td>
<td>Performance of DLS in detecting 14 different retinal conditions reported</td>
</tr>
<tr>
<td>Ming et al(^{62})</td>
<td>China</td>
<td>EyeWisdom</td>
<td>Prospective</td>
<td>199 participants</td>
<td>Two ophthalmologists</td>
<td>ICDRS</td>
<td>84.6% (54.6 to 98.1)</td>
<td>98.0% (94.3 to 99.6)</td>
<td>0.913 (0.797 to 1.00)</td>
<td>More than mild NPDR (macular signs not specified)</td>
<td>None</td>
</tr>
<tr>
<td>Pei et al(^{63})</td>
<td>China</td>
<td>EyeWisdom</td>
<td>Prospective</td>
<td>549 participants</td>
<td>Two ophthalmologists</td>
<td>ICDRS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>Sensitivity, specificity and AUC provided for each level of DR</td>
<td>None</td>
</tr>
<tr>
<td>Hao et al(^{64})</td>
<td>China</td>
<td>EyeWisdom</td>
<td>Prospective</td>
<td>3796 participants</td>
<td>Two attending physicians</td>
<td>ICDRS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>Reported sensitivity and specificity data for the presence of DR</td>
<td>None</td>
</tr>
<tr>
<td>Bawankar et al(^{65})</td>
<td>India</td>
<td>Bosch DR algorithm</td>
<td>Prospective</td>
<td>564 participants</td>
<td>The investigators</td>
<td>American Academy guidelines</td>
<td>91.18%</td>
<td>96.9%</td>
<td>NS</td>
<td>Figures refer to presence or absence of DR (macular signs not specified)</td>
<td>Bosch algorithm showed favorable performance for non-mydriatic photos</td>
</tr>
<tr>
<td>Study</td>
<td>Country</td>
<td>AI system name/developer</td>
<td>Study type</td>
<td>Number of participants or images</td>
<td>Who provided the reference standard</td>
<td>DR classification system used</td>
<td>Sensitivity (95% CI) for referable DR</td>
<td>Specificity (95% CI) for referable DR</td>
<td>AUC (95% CI) for referable DR</td>
<td>Referable DR threshold</td>
<td>Other reported outcomes</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>--------------------------</td>
<td>---------------------</td>
<td>----------------------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------</td>
<td>--------------------------------------</td>
<td>--------------------------------------</td>
<td>--------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hansen et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>Kenya</td>
<td>Iowa DR detection software</td>
<td>Retrospective</td>
<td>3460 participants</td>
<td>Moorfields Eye Hospital retinal image reading center</td>
<td>ICDRS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>Reported figures for the detection of the presence of DR</td>
</tr>
<tr>
<td>Dong et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>China</td>
<td>Retinal Artificial Intelligence Diagnosis System</td>
<td>Retrospective</td>
<td>110784 participants</td>
<td>Two ophthalmologists</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>0.981 (0.980 to 0.982)</td>
<td>NS</td>
<td>The AI system detects 10 retinal disorders</td>
</tr>
<tr>
<td>Yang et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>China</td>
<td>AIIDRScreening V.1.0</td>
<td>Prospective</td>
<td>962 participants</td>
<td>Two graders</td>
<td>Chinese Ophthalmic Society guidelines</td>
<td>86.72% (83.39 to 90.05)</td>
<td>96.06% (94.14 to 97.54)</td>
<td>NS</td>
<td>NS</td>
<td>Stage II or worse using the Chinese Ophthalmic Society guidelines</td>
</tr>
<tr>
<td>Mathenge et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>Rwanda</td>
<td>VISUHEALTH-AI DR (software V.1.8)</td>
<td>Prospective</td>
<td>31608 participants</td>
<td>Two optometrists or ophthalmologists</td>
<td>ICDRS</td>
<td>72.08% (70.68 to 73.46)</td>
<td>85.65% (85.11 to 86.18)</td>
<td>NS</td>
<td>NS</td>
<td>Moderate NPDR or worse and/or presence of DME</td>
</tr>
<tr>
<td>Nunez do Rio et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>India</td>
<td>VISUHEALTH-AI DR (software V.1.8)</td>
<td>Prospective</td>
<td>31608 participants</td>
<td>Two optometrists or ophthalmologists</td>
<td>ICDRS</td>
<td>72.08% (70.68 to 73.46)</td>
<td>85.65% (85.11 to 86.18)</td>
<td>NS</td>
<td>NS</td>
<td>Assessed AI system performance in non-mydriatic images</td>
</tr>
<tr>
<td>Li et al&lt;sup&gt;6&lt;/sup&gt;</td>
<td>China</td>
<td>DR DL Network (V.20171024)</td>
<td>Prospective</td>
<td>8012 participants</td>
<td>Two graders in reading center</td>
<td>ICDRS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>UK NSC guideline used to determine referral threshold (includes presence of DME) Al took 19.2 min to provide screening report after image capture vs 1.3 days for manual graders</td>
</tr>
</tbody>
</table>

*Study conducted in two sites; data presented separately for the two study sites.
†High-sensitivity cut-off data presented.
‡AUC presented from three different external testing datasets.
DR, detectable retinopathy; AUC, area under the curve; DLA, Deep learning algorithm; DL, deep learning system; DME, diabetic macular edema; DR, diabetic retinopathy; ICDRS, International Classification of Diabetic Retinopathy Severity; N/A, not available; NPDR, Non-proliferative diabetic retinopathy; NS, not stated; NSC, National Screening Committee; RDR, referable diabetic retinopathy.
sensitivity of 93.2% (95% CI 91.1% to 99.8%) and specificity of 88.7% (95% CI 88.3% to 89.0%) for referable AMD. The Comprehensive AI Retinal Expert system is a DLS designed to detect 14 retinal abnormalities from fundus imaging (including DR). The mean AUC for the detection of the 14 retinal pathologies in the three external test sets was 0.940 (SD 0.035), 0.965 (SD 0.031) and 0.983 (SD 0.042). This ranged from 0.861 (95% CI 0.788 to 0.922) for referable hypertensive retinopathy to 0.999 (95% CI 0.999 to 1.000) for geographic atrophy and retinitis pigmentosa; the AUC for referable DR in the non-Chinese ethnicity external dataset was 0.960 (95% CI 0.953 to 0.966).

Other outcomes reported included a pragmatic comparison of Google’s AI to local Thai graders. The AI had a sensitivity of 0.968 (range: 0.893–0.993), specificity of 0.956 (range: 0.983–0.987) and an AUC of 0.987 (range: 0.977–0.995), compared with a sensitivity of 0.734 (range: 0.4071–0.914) and a specificity of 0.980 (range: 0.939–1.000) for the regional graders; this difference was statistically significant (p<0.001). Another study from the Google health team reported the performance of a DLS in predicting macular edema from two-dimensional retinal photographs with a sensitivity of 81% and a specificity of 80%.

Some studies reported on the efficiency gains achieved when using AI-supported fundus image grading, highlighting the fact that patients received their screening result much more quickly when using AI compared with human graders.

**DISCUSSION**

There is considerable potential for AI to improve health services, particularly in LMICs. Ophthalmology is a potential exemplar medical specialty for healthcare AI, with its use in DR most advanced. We have identified 81 studies detailing the use of AI tools in LMICs in the context of DR. Over half of these report the use of AI to automate retinal image grading for DR.

Of the studies identified in this review, 43 were considered external validations. Thirty-one of those had a named AI and/or a stated developer. The reported sensitivities and specificities of these AI tools ranged from 83.3%–100% to 68.8%–98%, respectively providing evidence to support use in clinical practice. Google’s AI software, SELENA+ (EyRIS), EyeWisdom (Visionary Intelligence) and Medios AI (Remidio Innovation Solutions) accounted for about half of these publications and 13 were undertaken in, or on data from, China.

However, the majority of these 31 studies excluded ungradable images from the analyses and those that did not reported that the AI models considered a higher number of images ungradable compared with the reference standard human gradings. This suggests that if AI
Emerging technologies, pharmacology and therapeutics

systems are used prospectively in active clinical pathways, when ungradable images cannot be excluded, performance is likely to be reduced. If AI tools consider a higher proportion of images as ungradable, which typically trigger a refer outcome, this could result in more false positive cases being referred to and attending ophthalmology clinics, which are already under-resourced in many LMICs.

AI tools with other functions, including predicting the risk of DR progression and the use of a single AI model to detect multiple retinal diseases, give an indication of potential future developments. However, it is less clear from the literature if and how such systems can integrate into clinical pathways and whether their use will translate into improved health outcomes.

The majority of studies identified were conducted in two countries: India and China. These two countries account for a substantial proportion of the global population and it is therefore not surprising that they are also responsible for a disproportionate amount of healthcare AI research for DR in LMICs. Additionally, China’s and India’s middle-income (as opposed to low-income) status and more developed technology sectors has meant that there is greater in-country expertise and technology infrastructure to facilitate the development and testing of healthcare AI.

However, if AI for DR and other conditions is to be implemented and used to reduce healthcare disparities, as is often suggested, it is essential that contextually relevant research around the use of AI in clinical pathways is done in less well-resourced regions of the world. Datasets from populations in LMICs that are used to train AI models are necessary to prevent what has been coined ‘health data poverty’, whereby populations in poorer regions of the world, as well as minority ethnic groups in high-income countries, are disadvantaged through a lack of training data from such populations.

If these issues are not considered and addressed, global health inequities could be further exaggerated with wealthier countries that have invested in healthcare AI having access to, and using, new technologies and poorer countries left behind.

A further consideration is that screening for DR is only one part of a larger program that is required to reduce avoidable sight loss from DR. Improved access to retinal laser treatments and antivascular endothelial growth factor drugs is required with adequately trained eye care staff more widely available to deliver these treatments, particularly in low-resource settings. Without good access to affordable treatments that can be delivered effectively, improved screening for DR will not reduce sight loss from the disease.

However, before any of this becomes reality, healthcare AI needs to be integrated and used in clinical pathways. The current literature around AI for DR in LMICs has largely focused on AI’s performance in terms of sensitivities and specificities and does not adequately address the complex process of integrating this new technology into clinical care—a process which is likely to be even more challenging in LMICs.

We identified only three studies that focused on the implementation of an AI tool into an active clinical pathway in an LMIC; potential benefits included improved rates of follow-up following a point-of-care referral decision, although several challenges were also highlighted. The majority of studies described the development of AI models and only just over half were considered external validations. Of those AI tools that had been externally validated, we identified commercial websites for Medios AI, SELENA+, EyeArt, RAIDS and EyeWisdom and a non-commercial (charity) website for Cybersight, suggesting only some of the identified AI tools are ready for clinical deployment.

Few studies stated whether their AI tool had any regulatory approvals (eg, FDA or CE marking) and there is almost no available information on the cost of such systems, either in the literature or on commercial websites. These are all critical factors when hospitals and/or policy makers are deciding on whether to use AI in clinical care.

Moreover, a likely major advantage of healthcare AI for LMICs, as well as high-income countries, will be the potential health economic gains. We identified two economic evaluation studies, one of which concluded that AI for DR screening in Brazil was not cost-effective. The lack of transparency around the cost of AI systems makes such analyses difficult.

As we have highlighted, Google has a large portfolio of research around using AI for DR in LMICs, including one of only three studies that looked specifically at the implementation of AI for DR in an active clinical pathway. This paper candidly described the difficulties the team had and highlights the importance of implementation research embedded within prospective studies.

The requirement for a good internet connection to run their AI model was particularly highlighted as impractical. Other countries considering using AI that do not have access to a reliable and fast internet connection may face similar difficulties, suggesting models that can run on isolated machines offline may be more appropriate.

Despite Google’s large portfolio of research including implementation considerations, it should be noted that all of Google’s published work in Thailand and India was funded by Google with the majority of authors either employed by, or consultants for, Google. Additionally, there is no indication in any publication that Google’s AI system is or will be made available for use in clinical practice. If indeed the research is done with the intention of improving eye care, more transparency about access to and use of Google’s technology would be welcome.

In addition to the aforementioned clinical and implementation challenges, there are myriad legal and ethical issues adding further complexity which have not been addressed. For example, there are questions around accountability if errors are made, and legal frameworks for managing patient imaging and clinical data are
needed, although ethical considerations were beyond the scope of this review.

Investment in hardware infrastructure in LMICs that would enable patient data to be hosted on servers in the country where the AI is being used, for example, would provide a higher degree of control over data to the institutions and host countries, and would help with the curation of locally representative datasets, thereby addressing the issue of ‘health data poverty.’ Research programs investing in healthcare AI in LMICs have an opportunity to contribute to this, especially if work is done in collaboration with Ministries of Health.

The performance of the AI systems identified in this review demonstrates the potential for AI to improve diabetic eye care services in LMICs. There is a real opportunity for the quality of health service delivery in LMICs to be rapidly improved through leveraging such technologies. However, further research simply publishing the performance of AI tools in terms of sensitivities and specificities will not help this become reality.

The focus needs to move towards integrating AI models into health systems and detailing if and how their use improves clinical practice. Of the 81 studies included as full texts in this review, we identified only 1 randomized controlled trial. As AI tools are medical devices, it is important that, where possible, there is prospective clinical trial evidence to measure the effect of AI on clinical care prior to wider implementation. Clearer reporting of the impact of ungradable images on AI performance would also improve the evidence base.

Implementation research investigating how such systems can most effectively integrate into clinical pathways is needed as well as qualitative research specifically around acceptability and fidelity and LMIC population-specific dataset curation. As the two economic evaluations identified in this review demonstrate it is unclear whether using AI for DR screening is cost-effective, further work is needed in this regard.

The primary focus of the majority of studies identified in this review was sensitivity and specificity of the respective AI system to grade DR. However, no publication directly compared more than one AI model, therefore making it very difficult to compare the performance of different AI tools. Future validation work directly comparing different AI systems on the same image dataset by independent investigators would be of significant value and would enable a better comparison of performance. However, any commercially available AI systems included in such work should not be anonymized, otherwise comparative performance data would be of limited value.

If all this can be done, LMICs will be better placed to benefit from ongoing healthcare technology developments and, through the curation of LMIC population-specific datasets, will be able to maximize the performance of AI models in their populations. The potential of healthcare AI for DR as well as other conditions is arguably greatest in poorer regions of the world where there are fewer clinicians; however, there are a number of challenges to overcome if this potential is to be translated into reality.

**Contributors** CRC, JRE, IG, MJB and CB were responsible for study conception and design. IG developed the literature searches. CRC, CB and JR reviewed the literature. CRC, JR and CB abstracted and verified the data. CRC, CB, DML and MJB analyzed and interpreted the data. All authors had full access to all the data in the study. CRC drafted the original manuscript, and all coauthors reviewed the draft and provided critical feedback. All authors contributed to and approved the final manuscript.

**Funding** CRC is supported by the British Council for the Prevention of Blindness and the Sir Halley Stewart Trust. MJB is supported by the Wellcome Trust (207472/Z/17/Z).

**Disclaimer** No funders had any role in the writing of the manuscript or the decision to submit for publication.

**Map disclaimer** The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data sharing not applicable as no datasets generated and/or analyzed for this study.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

**ORCID iD**

Charles R Cleland http://orcid.org/0000-0002-2761-8092

**REFERENCES**


Emerging technologies, pharmacology and therapeutics


60 Re mendio Innovative Solutions Pvt Ltd. Available: https://www.re mendio.com/products/medios-dr
61 EYRIS Pte Ltd. Available: https://www.eyris.io
63 Sight AI Technology Pte Ltd. Available: https://www.sightai.com
64 Visionary intelligence (Vistel), n.d. Available: https://www.vistel.cn