Diagnostic accuracy of teleretinal screening for detection of diabetic retinopathy and age-related macular degeneration: a systematic review and meta-analysis

Parsa Mehraban Far,1 Felicia Tai,2 Adeteju Ogunbameru,3,4 Petros Pechlivanoglou,4,5 Beate Sander,3,4 David T Wong,2,6 Michael H Brent,2,7 Tina Felfeli2,3,4

Key messages

What is already known about this subject?
► With continuous advances in telecommunication technology and ophthalmic imaging in the last decade, teleretinal imaging is being relied on to identify patients with sight-threatening diabetic retinopathy (DR) and age-related macular degeneration (AMD).

What are the new findings?
► This meta-analysis of diagnostic test accuracy validates teleretinal screening as a highly accurate modality for diagnosis of any or referable DR, while evidence for diagnosis of AMD is more limited.

How might these results change the focus of research or clinical practice?
► More research is required to determine the diagnostic accuracy of teleretinal screening for diagnosis of AMD.
► The role of teleretinal screening relying on artificial intelligence should be examined in future research.

INTRODUCTION

Diabetic retinopathy (DR) and age-related macular degeneration (AMD) are among the leading causes of vision impairment in both low-income and high-income countries.1,2 Despite the significant advancements in therapeutics for both disorders, timely diagnosis and monitoring is essential for the prevention of irreversible vision loss.3 Traditional office-based face-to-face examination is effective for screening patients, but there are associated challenges in regions with limited accessibility to resources and eyecare specialists.4 Over the past decade with technological improvements, teleretinal screening has been explored as a cost-effective strategy to meet the increasing needs of the population worldwide.5,6

To date, a few systematic reviews have assessed the accuracy of teleretinal screening using human graders; however, there has been a limited number of meta-analyses with the use of correct hierarchical models to quantitatively summarise these results.7,8 Hierarchical methods are recommended as
The primary objective of this review was to assess the accuracy (sensitivity and specificity) of teleretinal screening for detection of DR and AMD compared with face-to-face clinical examination as a real-world reference standard. DR and AMD were chosen as the target conditions given that they account for the majority of cases in retina practices and have overlapping pathogenesis. The secondary objective was to formally assess the influence of exclusion of ungradable images in diagnostic accuracy calculations. This study adhered to the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA).

Search strategy
A comprehensive search of the literature was conducted on CENTRAL, Ovid MEDLINE and OVID Embase databases from 1 January 2010 to 25 July 2021. Detailed search strategy is included in online supplemental table 1. The time restriction for the search was placed to capture the most recent technological advances in teleretinal imaging modalities. No further restrictions were placed based on location, type and language of the publications. Retrieved studies were imported into Covidence (Melbourne, Australia), where duplicates were removed and article screening was performed.

Disease definition
Referrable DR was defined as any severity equivalent or worse than moderate non-proliferative DR or diabetic macular oedema (DMO). DMO was defined as any retinal thickening or the presence of hard exudate in the macula. Referrable AMD was defined as disease with features suggestive of intermediate and advanced state such as extensive intermediate drusen (<125 μm), any large drusen (>125 μm), neovascularisation or geographic atrophy.

Reference standard
The reference standard for determining diagnostic accuracy of teleretinal screening was chosen to be face-to-face clinical examination as opposed to the seven-field Early Treatment of Diabetic Retinopathy Study (ETDRS) protocol. The selection of this reference standard was based on three important justifications. First, face-to-face clinical examination is an established real-world modality for diagnosis and monitoring of patients with DR. Second, in-person examination is in keeping with the clinical pathway of most teleretinal screening programmes where selected patients with high-risk features or ungradable patients are ultimately referred for office-based examination. Lastly, face-to-face examination is a reputable consistent reference standard for all of the diagnoses included in the review including AMD, whereas the evidence supporting the ETDRS protocol stems from DR literature.

Study selection criteria
Two reviewers (PMF and FT) independently assessed eligibility first by the title and abstracts and then the full text of the retrieved studies. Conflicts not resolved were discussed with a third reviewer (TF) to reach consensus. Eligible studies included all comparative studies assessing the diagnostic accuracy using any form of retinal imaging modality for DR, DMO and AMD, where the reference standard was face-to-face examination using dilated funduscopy using direct/indirect ophthalmoscopy or slit lamp biomicroscopy. Retinal imaging modalities were defined broadly to include any form of fundus imaging device including handheld and table-top instruments regardless of the quality, field of view and wavelength of light used to capture the image.

Studies were excluded if they met any of the following exclusion criteria: (1) no reference standard of face-to-face ophthalmic examination, (2) no full-text available or insufficient information to allow for independent calculation of sensitivity and specificity and (3) grading of images not performed by human graders with specialty in ocular health (optometry/ophthalmology/special training attendee).

Data extraction and risk of bias assessment
Two independent reviewers (PMF, FT) were responsible for the extraction of relevant data from all included studies and assessing their methodological quality. Extracted data included: authors; year of publication; country or countries where the study was done; imaging devices used; imaging protocol and credential of image graders. For determination of diagnostic test accuracy, a two-by-two table was generated and the values corresponding to true positive, false positive, true negative and false negative were extracted for each calculation. In cases where information was insufficient, the data were requested via email from the corresponding authors of the publications. The QUADAS-2 tool was used independently by two reviewers (PMF, FT) to assess the methodological validity and applicability of each included study.

Data synthesis
The unit of analysis was number of eyes; however, number of patients was also accepted for analysis if
number of eyes screened was not reported. In cases where data were reported as patients only, each patient was counted as one unit of analysis. A sensitivity analysis was planned to assess the influence of using patients or eyes as the unit of analysis in the final diagnostic accuracy calculations. The random effects bivariate binomial model in R using the lme4 package was used to perform a meta-analysis and generate sensitivity, specificity, likelihood ratios (LR) and diagnostic OR associated with each test.\textsuperscript{22 23} The random effects bivariate binomial model was selected as it is a hierarchical method and is suitable for deriving summary sensitivity/specificity data at a specific threshold.\textsuperscript{24} Further details with regard to statistical analysis can be found in online supplemental appendix 1.

For the primary analysis, ungradable images were classified as having the target condition being assessed to simulate real-life patient care in a screening programme where undetermined results are further analysed and assumed positive. A sensitivity analysis was planned in advance, where ungradable images were excluded from analysis and diagnostic accuracy was calculated for only gradable images. A p value lower than 0.05 was used as the threshold for statistical significance. Summary estimates from statistical analyses were presented with their respective 95\% CIs along with p values where applicable.

Quality of evidence
Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach was used to assess the quality of evidence.\textsuperscript{25} Summary of findings tables along with GRADE evidence profiles were generated using GRADEpro software.\textsuperscript{26}

RESULTS
A total of 28 articles met the inclusion criteria for qualitative and quantitative synthesis. PRISMA flowchart of study selection process is depicted in figure 1. The most prevalent geographical location where the studies were conducted was the USA accounting for 36\% (10/28), followed by Europe 25\% (7/28). For studies reporting on patients with DR, the mean number of years from initial diagnosis of diabetes was greater than or equal to 5 years for all studies at the time of publication where this was reported. The pooled cohort included both type I and II diabetes; however, the relative proportion of each type was not consistently presented in the included studies. All studies, except for 11\% (3/28) used table-top fundus cameras for screening. The reference standard of choice was dilated fundus examination with slit-lamp biomicroscopy or binocular indirect ophthalmoscopy for 82\% (23/28) of studies. Descriptive details of the eligible studies are provided in online supplemental table 2.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{Preferred Reporting Items for a Systematic Review and Meta-analysis flow diagram for transparent reporting of study selection process and meta-analysis.}
\end{figure}
Additional information such as diagnostic criteria for referrable disease in each study as well as the number and process of grading are presented in online supplemental tables 3 and 4.

**Diagnostic accuracy for detectable DR**

Thirteen studies including 7207 eyes contributed to the meta-analysis (figure 2). Teleretinal screening was found to have a specificity of 0.88 (95% CI: 0.74 to 0.95) and sensitivity of 0.91 (95% CI: 0.82 to 0.96) for detection of DR. Diagnostic OR, LR+ and LR− associated with the accuracy of fundus imaging were calculated to be 77.59 (95% CI: 29.88 to 201.50), 7.78 (95% CI: 3.43 to 17.65) and 0.1 (95% CI: 0.05 to 0.20), respectively. There was a large degree of heterogeneity in the results from individual studies corresponding to a large predictive region in the summary receiver operating characteristic (sROC) curve (online supplemental figure 1).

Investigation of heterogeneity through meta-regression was only possible for mydriasis due to paucity of data for other covariates. Both sensitivity and specificity were higher in dilated eyes in comparison to undilated eyes with a sensitivity of 0.91 (95% CI: 0.82 to 0.95) versus 0.89 (95% CI: 0.66 to 0.97) and specificity of 0.89 (95% CI: 0.63 to 0.97) versus 0.85 (95% CI: 0.71 to 0.93), respectively. Addition of pupil status as a covariate for both sensitivity and specificity to the model did not achieve statistical significance ($\chi^2(5)=10.31, p=0.07$).

**Diagnostic accuracy for AMD**

Due to paucity of data, the bivariate binomial model could not be fitted for analysis. A univariate random effects model was fitted by removing the correlation term between sensitivity and specificity. The raw diagnostic accuracy values for each individual study are provided in online supplemental table 5. Three studies including 697 eyes contributed to the univariate random effects meta-analysis for detection of any AMD with an overall sensitivity of 0.71 (95% CI: 0.49 to 0.86) and specificity of 0.88 (95% CI: 0.85 to 0.90).

**Use of ancillary imaging**

Only one study assessed the influence of addition of optical coherence tomography (OCT) to a teleretinal screening programme.17 The authors concluded that OCT did not improve the diagnostic accuracy of their teleophthalmology programme for the detection of glaucoma, DR and AMD.

**Methodological quality**

Online supplemental figure 4 depicts a graphical representation of summary quality and applicability to our research question of the included studies. Selection bias was noted based on exclusion of some data for patients with corneal disorders, media opacity and ungradable images from the analysis in some studies. Online meta-regression could be performed due to paucity of data. There was a mild-moderate degree of heterogeneity as depicted in the sROC curve in online supplemental figure 3.

**Diagnostic accuracy for referable AMD**

No meta-analysis for referable AMD was performed due to significant interstudy variability in the definition of referable AMD.
supplemental figure 5 depicts the detailed assessment of quality in each individual study.

Strength of evidence
Overall, quality of body of evidence was low to moderate for the reported outcomes. Summary of findings tables and GRADE profiles are demonstrated in online supplemental table 6.

Sensitivity analyses
Diagnostic accuracy for DMO
A sensitivity analysis was performed to quantify the accuracy of teleretinal screening for identifying DMO. Six studies including 4255 eyes contributed to the meta-analysis (figure 3). Teleretinal screening maintained a favourable diagnostic accuracy, but lower than that of any DR, with a sensitivity of 0.84 (95% CI: 0.76 to 0.90) and specificity of 0.85 (95% CI: 0.75 to 0.91) for detection of DMO. Diagnostic OR, LR+ and LR− were calculated to be 30.57 (95% CI: 15.20 to 61.48), 5.61 (95% CI: 3.39 to 9.28) and 0.18 (95% CI: 0.12 to 0.29), respectively. No formal meta-regression could be performed due to the limited number of studies contributing to this analysis. There was a mild-moderate degree of heterogeneity with a relatively small predictive region in sROC curve (online supplemental figure 6).

Exclusion of ungradable images
Twelve studies including 8452 eyes contributed to the meta-analysis for detection of any level of DR (online supplemental figure 7). After the exclusion of ungradable images, diagnostic accuracy for detection of any levels DR remained high with a sensitivity of 0.88 (95% CI: 0.75 to 0.94) and specificity of 0.90 (95% CI: 0.81 to 0.96). A total of 13 studies, including 6481 eyes, were analysed for meta-analysis of referrable DR (online supplemental figure 8). After exclusion of ungradable cases, the specificity increased to 0.95 (95% CI: 0.90 to 0.98), while the sensitivity remained nearly unchanged at 0.85 (95% CI: 0.76 to 0.91).

Patients and eyes as unit of analysis
Separate meta-analysis in studies which used patients versus eyes as unit of analysis for detection of any level of DR was performed. Meta-analysis using eyes as unit of analysis for detection of any level of DR was performed using data from eight studies including 4299 eyes. Teleretinal screening had a sensitivity of 0.95 (95% CI: 0.87 to 0.98) and specificity of 0.88 (95% CI: 0.70 to 0.96). Using patients as the unit of analysis based on five studies and 2908 patients, teleretinal screening achieved a sensitivity of 0.80 (95% CI: 0.59 to 0.91) and specificity of 0.88 (95% CI: 0.60 to 0.98).

DISCUSSION
Here, we presented our findings from a systematic review and meta-analysis on estimates of sensitivity and specificity of teleretinal screening for detection of DR and AMD when compared with face-to-face clinical examination as the real-world reference standard. We found that teleretinal screening achieved a high accuracy for detection of any DR with a sensitivity of 0.91 (95% CI: 0.80 to 0.95) and specificity of 0.95 (95% CI: 0.93 to 0.98). The diagnostic accuracy teleretinal screening has been previously characterised at specific levels of DR severity; however, data on the overall accuracy for detection of referable cases have not been consistently reported.

Figure 3  Forest plots depicting sensitivity and specificity of teleretinal screening for detection of diabetic macular oedema.
of accuracy for detection of DMO and referrable DR in comparison to any DR.

Given that the previous reviews to date on this topic have typically excluded ungradable images from their analysis, our sensitivity analysis with the exclusion of ungradable images showcases a cautionary message for future investigators. In fact, the specificity of fundus imaging for identification of referral-warranted DR improved by nearly 10% after ungradable images were removed from analysis. This observation is expected and can be explained by spectrum effect, whereby systematic removal of a patient subgroup, such as difficult to diagnose cases with media or corneal opacity, leads to an easier diagnosis and detection of referable and non-referrable cases.

Based on our findings, evidence in support of implementation of teleretinal screening for detection of AMD was limited in comparison to DR. Only three studies provided diagnostic accuracy data for detection of any AMD. Although these results are encouraging, with an overall sensitivity of 0.71, more research is required to establish a role for fundus imaging for diagnosis and treatment of patients with AMD. One strategy to generate more diagnostic accuracy data for AMD detection is to implement AMD detection into the already existing teleretinal screening infrastructure for DR. If teleretinal screening prove to be a highly accurate tool within the structure of the pre-existing teleophthalmology programmes, further emphasis may be placed on detection of AMD.

The role of OCT for diagnosis and monitoring of retinal disorders is well established. However, whether its incorporation into teleophthalmology screening programmes is beneficial remains controversial. Only one study in this systematic review provided a direct comparison in the diagnostic accuracy of fundus photography combined with OCT and fundus photography alone. In the paper by Maa et al, despite the detailed cross-sectional analysis of the macula, optic nerve head and retinal nerve fibre layer that is provided in OCT scans, the authors did not detect any improvement in sensitivity or specificity of fundus photography for the detection of glaucoma and retinal disorders with the addition of OCT. In contrast to these results, other groups have clearly demonstrated a role for OCT in addition to fundus photographs, especially for the detection of diabetic macula oedema which requires a stereoscopic view. In the current meta-analysis, only a fraction of the studies used OCT alone or as an adjunct modality in addition to fundus photographs and we were unable to perform a formal meta-regression. Due to widespread use of OCT as well as recent advances in OCT technology such as swept-source OCT and OCT angiography, it is inevitable that more data will become available in the near future.

Meta-regression based on pupil status showed a sizeable improvement in diagnostic accuracy when eyes were dilated prior to capturing of the image which approached statistical significance. Although we are unable to identify the exact reason behind this observation, it can be hypothesised that a larger pupil diameter allows for increased capture of light by the camera and therefore generates a higher quality image. This finding should be verified in different ethnic groups as well as in individuals with difference in iris colour which could elicit different levels of response to pharmacological dilation.

**Limitations and future directions**

It is important to note that there is a large degree of heterogeneity in the diagnostic criteria for referrable DR and AMD. Additionally, there is also a large degree of heterogeneity in the sample including patients with type I and II diabetes of differing durations. Similar to all review papers, publication bias may be present whereby studies that achieve high diagnostic accuracy are preferentially published in comparison to studies where the accuracy is lower which could lead to an overestimation of sensitivity and specificity. Our study results are only applicable to teleretinal programmes using human graders. Recent diagnostic test accuracy meta-analyses have provided very promising accuracy estimates for machine-learning-based teleretinal screening programmes for DR. Future studies should assess the diagnostic accuracy of automated systems using artificial intelligence and deep-learning algorithms in teleophthalmology screening programmes for ocular diseases. Lastly, the focus of this review was on teleretinal screening for the most common retinal pathologies. As more data become available, future investigations should assess the utility of teleglaucoma screening programmes.

**Author affiliations**

1Department of Ophthalmology, Queen’s University, Kingston, Ontario, Canada
2Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada
3Toronto Health Economics and Technology Assessment (THETA) Collaborative, University Health Network, Toronto, Ontario, Canada
4Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
5Peter Gilgan Centre for Research and Learning, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada
6Department of Ophthalmology, St. Michael’s Hospital, Toronto Unity Health, Toronto, Toronto, Ontario, Canada
7Retina Service, Donald K. Johnson Eye Institute, University Health Network, Toronto, Ontario, Canada

**Twitter** Tina Felfeli @TinaFelfeli

**Acknowledgements** We would like to acknowledge the contributions of the Diabetes Action Canada, a Canadian Institutes for Health Research (CIHR) Strategy for Patient-Oriented Research Network in Chronic Disease in supporting this research project.

**Contributors** PMF, TF and BS designed the study. PMF, FT and TF screened papers and extracted data. PMF, TF, PP and AO analysed data and provided statistical and methodological support. All authors interpreted data and prepared the manuscript. TF is responsible for the overall content as the guarantor.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** This study does not involve human participants.

**Provenance and peer review** Not commissioned; externally peer reviewed.
26 GRADEPro G. Computer program. McMaster University (developed by evidence prime). GRADEPro GMT. Hamilton (ON): McMaster University (developed by Evidence Prime), 2015.