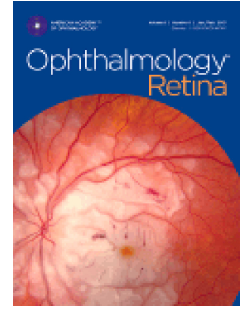


Journal Pre-proof



Deep learning to detect optical coherence tomography-derived diabetic macular edema from retinal photographs: a multicenter validation study

Xinle Liu, PhD, Tayyeba K. Ali, MD, MS Preeti Singh, Ami Shah, MD, Scott Mayer McKinney, MS, Paisan Ruamviboonsuk, MD, Angus W. Turner, MSc, FRANZCO, Pearse A. Keane, MD, FRCOphth, Peranut Chotcomwongse, MD, Variya Nganthavee, MD, Mark Chia, MBBS, MMed, Josef Huemer, MD, FEBO, Jorge Cuadros, OD, PhD, Rajiv Raman, MBBS, MS, Greg S. Corrado, PhD, Lily Peng, MD, PhD, Dale R. Webster, PhD, Naama Hammel, MD, Avinash V. Varadarajan, MS, Yun Liu, PhD, Reena Chopra, BSc, Pinal Bavishi, BE

PII: S2468-6530(22)00001-X

DOI: <https://doi.org/10.1016/j.oret.2021.12.021>

Reference: ORET 1142

To appear in: *Ophthalmology Retina*

Received Date: 13 September 2021

Revised Date: 9 November 2021

Accepted Date: 29 December 2021

Please cite this article as: Liu X., Ali T.K., Singh P., Shah A., McKinney S.M., Ruamviboonsuk P., Turner A.W., Keane P.A., Chotcomwongse P., Nganthavee V., Chia M., Huemer J., Cuadros J., Raman R., Corrado G.S., Peng L., Webster D.R., Hammel N., Varadarajan A.V., Liu Y., Chopra R. & Bavishi P., Deep learning to detect optical coherence tomography-derived diabetic macular edema from retinal photographs: a multicenter validation study, *Ophthalmology Retina* (2022), doi: <https://doi.org/10.1016/j.oret.2021.12.021>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1 Deep learning to detect optical coherence tomography-
2 derived diabetic macular edema from retinal
3 photographs: a multicenter validation study

4
5 Xinle Liu, PhD^{1†}, Tayyeba K. Ali, MD^{2,3†}, Preeti Singh, MS¹, Ami Shah, MD², Scott Mayer
6 McKinney, MS¹, Paisan Ruamviboonsuk, MD⁴, Angus W. Turner, MSc, FRANZCO^{5,6}, Pearse
7 A. Keane, MD, FRCOphth⁷, Peranut Chotcomwongse, MD⁴, Variya Nganthavee, MD⁴, Mark
8 Chia, MBBS, MMed^{5,7}, Josef Huemer, MD, FEBO⁷, Jorge Cuadros, OD, PhD⁸, Rajiv Raman,
9 MBBS, MS⁹, Greg S. Corrado, PhD¹, Lily Peng, MD, PhD¹, Dale R. Webster, PhD¹, Naama
10 Hammel, MD^{1*}, Avinash V. Varadarajan, MS¹, Yun Liu, PhD¹, Reena Chopra, BSc^{1,7*}, Pinal
11 Bavishi, BE¹

12
13 ¹ Google Health, Google LLC, Mountain View, CA, USA

14 ² Work done at Google Health via Advanced Clinical, Deerfield, IL, USA

15 ³ Present affiliation: California Pacific Medical Center, Department of Ophthalmology, San
16 Francisco, CA

17 ⁴ Department of Ophthalmology, Rajavithi Hospital, College of Medicine, Rangsit University,
18 Bangkok, Thailand

19 ⁵ Lions Outback Vision, Lions Eye Institute, Nedlands, Western Australia, Australia

20 ⁶ University of Western Australia, Perth, Western Australia, Australia

21 ⁷ NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS
22 Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom

23 ⁸ EyePACS Inc., Santa Cruz, CA, USA

24 ⁹ Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai, India

25
26 [†] Equal contribution

27 * Co-corresponding authors

28

29 **Meeting presentation**

30 Part of this work has been presented as an abstract at ARVO Annual Meeting 2020.

31

32 **Financial support**

33 Google LLC funded this study, and participated in the design of the study, conducting the
34 study, data collection, data management, data analysis, interpretation of the data,
35 preparation, review and approval of the manuscript.

36

37 **Conflicts of interest**

38 P.R. is a consultant for Roche, Novartis, and Bayer. P.A.K. has acted as a consultant for
39 DeepMind, Roche, Novartis, and Apellis and is an equity owner in Big Picture Medical. He
40 has received speaker fees from Heidelberg Engineering, Topcon, Allergan, and Bayer. He is
41 supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK
42 Research & Innovation Future Leaders Fellowship (MR/T019050/1). J.H. served on advisory
43 boards for Roche. He has received speaker fees from Bayer and Zeiss. J.C. is CEO of
44 EyePACS, Inc.. X.L., P.S., S.M.M., G.S.C., L.P., D.R.W., N.H., A.V.V., Y.L., R.C., and P.B.
45 are employees of Google LLC and own Alphabet stock.

46

47 **Running head**

48 Deep learning to predict OCT-derived labels of DME from CFP

49

50 **Address for reprints**

51 Naama Hammel, MD, Google Health, Google LLC, 1600 Amphitheatre Parkway, Mountain
52 View, CA 94043, USA. nhammel@google.com

53 Reena Chopra, BSc, Google Health, Google LLC, 6 Pancras Square, Kings Cross, London
54 N1C 4AG, UK. reenac@google.com

55

56 **Keywords**

57 Diabetic Retinopathy, Diabetic Macular Edema, Deep Learning, Optical Coherence

58 Tomography, Artificial Intelligence

59

60 **Word count**

61 3848 words

62

63 **Acknowledgments**

64 The authors would like to thank Ehsan Rahimy, MD, Siva Balasubramanian, MD, PhD, Ilana

65 Traynis, MD, Ali Zaidi, MD, and all the graders for guideline iteration and labeling retinal

66 images; Mahvish Nagda, MS, Derek Wu, BS, Jieying Luo, PhD, Roy Lee, BS, Anita Misra B.

67 Tech, Richa Tiwari, PhD, Fred Hersch, MBBS, MPH, Rajroshan Sawhney, BE, Yan Zhen

68 Tan, LLB, Warisara Pattanapongpaiboon, MD, Vaibhav Shah, DNB, for technical and

69 logistical support; Divleen Jeji, MD, MBA and Craig Mermel, MD, PhD for their feedback on

70 the manuscript; and the Google Health labeling software infrastructure team for their

71 assistance.

72 Abstract

73

74 **Purpose**

75 To validate the generalizability of a deep learning system (DLS) that detects diabetic
76 macular edema (DME) from two-dimensional color fundus photography (CFP), where the
77 reference standard for retinal thickness and fluid presence is derived from three-dimensional
78 optical coherence tomography (OCT).

79

80 **Design**

81 Retrospective validation of a DLS across international datasets.

82

83 **Participants**

84 Paired CFP and OCT of patients from diabetic retinopathy (DR) screening programs or
85 retina clinics. The DLS was developed using datasets from Thailand, the United Kingdom
86 (UK) and the United States and validated using 3,060 unique eyes from 1,582 patients
87 across screening populations in Australia, India and Thailand. The DLS was separately
88 validated in 698 eyes from 537 screened patients in the UK with mild DR and suspicion of
89 DME based on CFP.

90

91 **Methods**

92 The DLS was trained using DME labels from OCT. Presence of DME was based on retinal
93 thickening or intraretinal fluid. The DLS's performance was compared to expert grades of
94 maculopathy and to a previous proof-of-concept version of the DLS. We further simulated
95 integration of the current DLS into an algorithm trained to detect DR from CFPs.

96

97 **Main Outcome Measures**

98 Superiority of specificity and non-inferiority of sensitivity of the DLS for the detection of
99 center-involving DME, using device specific thresholds, compared to experts.

100

101 **Results**

102 Primary analysis in a combined dataset spanning Australia, India, and Thailand showed the
103 DLS had 80% specificity and 81% sensitivity compared to expert graders who had 59%
104 specificity and 70% sensitivity. Relative to human experts, the DLS had significantly higher
105 specificity ($p=0.008$) and non-inferior sensitivity ($p<0.001$). In the UK dataset the DLS had a
106 specificity of 80% ($p<0.001$ for specificity $> 50\%$) and a sensitivity of 100% ($p=0.02$ for
107 sensitivity $> 90\%$).

108

109 **Conclusions**

110 The DLS can generalize to multiple international populations with an accuracy exceeding
111 experts. The clinical value of this DLS to reduce false positive referrals, thus decreasing the
112 burden on specialist eye care, warrants prospective evaluation.

113 Introduction

114 Diabetic macular edema (DME) is characterized by retinal thickening and an accumulation of
115 intraretinal fluid (IRF) caused by abnormal vascular permeability and leakage in diabetic
116 retinopathy and diabetes mellitus, and is a leading cause of blindness among working-aged
117 adults.^{1,2} Prompt detection and treatment of DME is imperative to stabilize vision.³

118

119 Diabetic eye screening programs utilize colour fundus photography (CFP) for detection of
120 diabetic retinopathy (DR) and DME. For DME, the presence of hard exudates (HE) near the
121 fovea is used as a surrogate marker for the presence of fluid. However, this marker alone is
122 an imperfect indicator of DME and has limited specificity and sensitivity.^{4,5,6}

123

124 Clinical testing via three-dimensional optical coherence tomography (OCT) is increasingly
125 recognized as the reference standard for detection of DME.⁷ Clinical trials investigating DME
126 therapies have relied on OCT-derived central retinal thickness as an inclusion criteria and
127 clinical endpoint.^{8,9} Another important marker of DME activity on OCT is the presence of
128 intraretinal fluid (IRF), which may gauge response to treatment.¹⁰ Despite the advantage of
129 OCT for reliable diagnosis and classification of DME, this modality remains unavailable in
130 many parts of the world due to its high cost and need for expert interpretation,¹¹ resulting in
131 ongoing reliance on CFP for DME screening.

132

133 Deep learning, a type of artificial intelligence, has been used for computational detection of
134 DR and DME from CFP.¹²⁻¹⁵ These deep learning systems (DLSs) are typically trained using
135 CFP with HE-labels and are unlikely to address the sensitivity and specificity gaps. To
136 improve upon expert performance, our group previously reported a DLS that predicted OCT-
137 derived DME labels using only CFP as input (the "DME-DLS").¹⁶ Other groups have similarly
138 evaluated feasibility of systems that predict OCT-derived central retinal thickness using a
139 CFP input and obtained promising results.¹⁷

140

141 This is the first study of its kind to generalize a DLS aimed at predicting OCT-derived ground
142 truth using CFP as input in multiple, independent screening populations. The performance of
143 the DME-DLS was compared to both expert grading of CFP and the previously reported
144 version of the DLS. Additionally, we analyzed the ability of the DLS in detecting center-
145 involving DME (ci-DME) with definitions based on retinal thickening and IRF presence. As a
146 secondary outcome, the DLS performance for detecting the presence of OCT-based DME
147 irrespective of location was evaluated. To better understand the potential impact of improved
148 DME screening in the context of broader DR screening, we further evaluated the net effect of
149 replacing the DME component of a previously published and extensively validated Krause et.
150 al DLS that detects both DR and DME using conventional grading guidelines.^{18,19,20}

151 Methods

152 Study data were obtained from multiple sources. These data were collected with the
153 participant's consent and/or de-identified in accordance with local regulatory requirements
154 (e.g., HIPAA) and/or reviewed by the institution's Ethics Committee or IRB prior to our
155 receipt of the dataset. The study adhered to the tenets of the Declaration of Helsinki.

156 Datasets

157 Development datasets

158 Our DLS was developed using 1,167,791 retrospectively collected paired single-field CFP
159 and OCT images from four sites in three countries: Thailand (1,299 images from Lerdsin
160 Hospital and 7,072 from Rajavithi Hospital), the UK (1,156,142 images from Moorfields Eye
161 Hospital), and the US (3,278 images from Alameda County Health System) (Table S1). All
162 images were collected from diabetic patients except for the UK dataset, which consisted of a
163 wide range of retinal pathology. Data were divided randomly, by patient, into train (98.8%)
164 and tune (1.2%) sets, with the ratio based on an empirical estimation of necessary tuning
165 versus training dataset sizes. The train set came from Thailand Lerdsin (100%), Thailand
166 Rajavithi (68%), and the UK (99.3%), while the tune set was gathered from Thailand
167 Rajavithi (32%), the UK (0.7%) and the US (100%). Some datasets were used exclusively
168 for training or tuning to help assess inter-dataset generalization during tuning, thus their
169 ability to generalize to new datasets.

170 Validation datasets

171 Validation of the DLS was performed on independent datasets, comprising patients with
172 diabetes from institutions in Australia, India, Thailand and the UK. Datasets from Australia
173 and India were external validation sets. The validation dataset from Thailand was from the
174 same institution as the developmental set; however, it was from a different temporal period

175 and consisted solely of a screening population (compared to retina clinic patients). The UK
176 dataset was an internal validation dataset from the same institution and an overlapping time
177 period (though without patient duplication) (Table 1).

178 Definitions, image acquisition, and grading

179 Definition of OCT-based DME

180 DLS performance was evaluated using two separate definitions of DME: one based on
181 retinal thickening and the second on IRF presence; each definition was further divided into
182 ci-DME and DME (combined ci-DME and non-center involving [nci-DME]).

183

184 For retinal thickening, device- and gender-specific thresholds were used to define ci-DME in
185 ETDRS zone 1/central subfield thickness (CST) in all datasets (Table S2).^{3,21–23} For IRF
186 presence, ci-DME was defined as fluid present within ETDRS zone 1. Similarly, nci-DME
187 was defined as retinal thickening or IRF occurring solely in zones 2–9.²⁴ The retinal
188 thickening based ci-DME definition was used for primary analysis; all other definitions were
189 used for secondary analysis. Further details of defining DME, OCT retinal thickness
190 acquisition and OCT grading of fluid presence can be found in the Supplementary Methods
191 S1.

192 CFP grading for comparison to DLS

193 To provide a baseline comparison for the DLS, experienced graders labeled CFP in the
194 validation set for DR severity and for the presence and location of HE within 1500 μm of the
195 foveal center. DR level was based on the International Clinical Diabetic Retinopathy Disease
196 Severity Scale²⁵ and presence of HE, reflecting clinical practice in DR screening programs.
197 Further details on grading CFPs and image quality guidelines are provided in the
198 Supplementary Methods S1.

199

200 Table 2 describes the availability of ground truths and expert grades for comparison in the
201 validation datasets.

202 Deep learning system

203 Development

204 The DME-DLS was trained similarly to the proof-of-concept version,¹⁶ with some significant
205 upgrades to aid generalization. The DME-DLS is a deep convolutional neural network
206 trained with TensorFlow.²⁷ The DLS takes CFP as input and simultaneously outputs
207 predictions for thickness-based ci-DME, IRF-based ci-DME, and thickness-based DME
208 (inclusive of both ci-DME and nci-DME). While OCT provided the ground truth labels, the
209 actual OCT images were never seen by the DLS during training or validation (Figure 1A). To
210 improve performance, the DLS was developed using multi-task learning with the following
211 co-trained tasks: subretinal fluid presence, prediction of sex, and prediction of age. To aid
212 generalization, this version of the DME-DLS was also trained on the large UK development
213 dataset, with automated labels generated using a previously described segmentation DLS.²⁸
214 The current model was trained to predict the volume of fluid output by the segmentation DLS
215 instead of fluid presence alone and it used the EfficientNet-B5 architecture²⁹ to train on the
216 much larger dataset in a reasonable amount of time. Further details on the DLS design are
217 available in Supplementary Methods S2.

218 Evaluation

219 To enable comparison with expert grades, the output of the DLS was thresholded to produce
220 a binary result: DME present or absent. As the Australia, India and Thailand datasets
221 consisted of a screening population where the pre-test probability for presence of DME
222 would be lower, high-specificity operating points were chosen, with separate operating points
223 for ci-DME (thickness and IRF presence) and DME (thickness) outputs. Operating points
224 were selected to maximize specificity on the Thailand and US tune sets, subject to the

225 constraint that the 95% confidence lower bound of DME-DLS sensitivity was at most 10%
226 lower than the expert graders. The same operating points were used for the aggregated
227 analysis of all three datasets as well as for each individual country's dataset. Since the UK
228 dataset contained data from a pre-screened population with evidence of mild DR and a
229 suspicion of DME, the pre-test probability for presence of DME would be higher, and so a
230 separate high-sensitivity operating point was chosen with 95% confidence lower bound
231 sensitivity of 90%. All operating points were pre-selected before evaluating the model on the
232 validation datasets (Figure 1B).

233

234 We further simulated integration of this DME-DLS into another previously validated DLS,¹⁸
235 with the originally described DLS being referred to as the Krause et al. DLS henceforth. The
236 Krause et al. DLS detects DR and DME and was developed on an independent dataset of
237 CFP using the presence of HE as labels for DME. Our DME-DLS was used to replace the
238 DME detector in the Krause et al. DLS without modifying the DR component (Figure S1).
239 Using the adjudicated Indian dataset, we then compared the original and modified version of
240 the Krause et al. DLS for detecting patients with vision-threatening DR (VTDR), defined as
241 severe non-proliferative DR, proliferative DR, or ci-DME (retinal thickening) in at least one
242 eye.

243 Statistical analysis

244 For pre-specified primary analyses, the performance of the DLS was compared with experts,
245 specifically testing superiority of specificity and non-inferiority of sensitivity (at a 10% margin)
246 for detection of ci-DME (thickening) on the aggregation of Australia, India, and Thailand
247 validation sets. Superiority comparisons for specificity were two-sided McNemar tests,³⁰
248 while non-inferiority comparisons for sensitivity with a pre-specified margin of 10% were
249 Wald tests.³¹ Since there were multiple graders, we adjusted for the clustered nature of the
250 data using Obuchowski's method for paired binomial proportions.^{32,33} Confidence intervals
251 for sensitivities and specificities were calculated using the exact Clopper-Pearson interval.

252 We further conducted receiver operating characteristic (ROC) analysis to evaluate the DLS
253 both in isolation and in comparison to the previously reported version of the DLS. Non-
254 parametric confidence intervals on the area under the curve (AUC)-ROC were computed
255 with DeLong's method.³⁴

256

257 Secondary analysis evaluated DLS performance when compared to experts on a per-dataset
258 level (Australia, India and Thailand). We also analyzed DLS performance for detection of
259 DME defined by fluid presence, DME irrespective of location, and when restricted to mild-to-
260 moderate DR, where ground truth was available. On the UK dataset, we tested for sensitivity
261 of > 90% and specificity of > 50% using an exact one-sided binomial test.

262 Results

263 The DME-DLS was validated on independent, screening datasets from Australia, India and
264 Thailand. Additionally, the DME-DLS was validated in a cohort of patients from the UK
265 screening program referred specifically for maculopathy based on CFP. The characteristics
266 of the cohorts are provided in Table 2. The rate of ci-DME was higher in the India dataset
267 (21%) compared to the other 3 datasets (3–5%). After excluding ungradable images, the
268 DLS was validated on 3574 images.

269 DME based on retinal thickening

270 Our primary analysis evaluated the DME-DLS against a reference standard of ci-DME based
271 on retinal thickening in a combined cohort of the Australia, India, and Thailand datasets and
272 compared it to experts grading maculopathy on CFP. The DME-DLS had a specificity of 80%
273 and a sensitivity of 81% compared to expert specificity of 59% and sensitivity of 70%. The
274 DME-DLS had superior specificity ($p=0.008$) and non-inferior sensitivity ($p<0.001$), reducing
275 false positives by 51%. Exploratory analysis revealed the DME-DLS also had superior
276 sensitivity ($p=0.014$) (Table 3).

277

278 Per-dataset, the DME-DLS's specificities (vs human experts) for the Australia, India, and
279 Thailand datasets were 86% (vs 57%; $p=0.03$), 77% (vs 44%; $p<0.001$) and 66% (vs 78%;
280 $p=0.20$), respectively; the corresponding sensitivities were 71% (vs 66%, $p=0.007$ for non-
281 inferiority), 84% (vs 72%, $p=0.002$) and 100% (vs 77%, $p=0.013$) (Table 3).

282

283 We also compared our DME-DLS's ci-DME detection performance with a previously
284 reported, non-generalized, proof-of-concept version of this DLS (Figure 2A).¹⁶ The current
285 DME-DLS had an AUC of 0.88 compared to 0.80 for the proof-of-concept DLS. For the
286 Australia, India, and Thailand datasets individually, the AUCs for the DME-DLS (vs the

287 proof-of-concept version) were 0.86 (vs 0.73), 0.89 (0.74), 0.96 (0.93), respectively. The
288 performance was consistent when restricted to eyes with mild-to-moderate DR (Figure S2)
289 and when subgrouped by gender (Figure S3).

290

291 For the UK dataset, the DME-DLS had an AUC of 0.96 (vs 0.82 for the proof-of-concept
292 DLS), a specificity of 88% and a sensitivity of 89%. At the prespecified high-sensitivity
293 threshold, the DME-DLS had a specificity of 80% ($p < 0.001$ for specificity $> 50\%$) and a
294 sensitivity of 100% ($p = 0.02$ for sensitivity $> 90\%$) (Figure 2B). The DME-DLS reduced false
295 positives by 80% from 661 to 129, while detecting all 37 eyes with ci-DME on OCT.

296

297 Besides ci-DME, we evaluated the DME-DLS against a reference standard of DME (based
298 on retinal thickening) anywhere in the macula (Figure S4). The DME-DLS specificities (vs
299 human experts) for the Australia, India and Thailand datasets were 93% (vs 69%), 87% (vs
300 55%) and 70% (vs 86%) respectively; the corresponding sensitivities were 56% (vs 60%),
301 76% (vs 72%) and 90% (vs 65%), respectively.

302 DME based on intraretinal fluid presence

303 When assessing DLS performance against the expert graded reference standard in
304 detecting DME defined by IRF, similar trends in performance were noted (Figure 2C-D). In
305 the Australia and India validation datasets, at the pre-specified threshold, the DLS
306 specificities (vs human experts) were 92% (vs 68%) and 67% (vs 44%), respectively; the
307 corresponding sensitivities were 64% (vs 63%) and 89% (vs 71%), respectively. Additionally,
308 the DLS had an AUC of 0.86 and 0.88 compared to 0.74 and 0.77, respectively, for the
309 previously reported proof-of-concept DLS (Figure 2C). There was no significant impact on
310 performance of the DLS when the reference standard was expanded to IRF presence
311 anywhere in the macula (Figure S3D).

312

313 For the UK dataset, the DLS had a sensitivity of 56% and a specificity of 95%. At the
314 prespecified high-sensitivity threshold the DLS had a sensitivity of 94% ($p=0.032$ for
315 sensitivity $> 90\%$) and a specificity of 52% ($p<0.001$ for specificity $> 45\%$) (Figure 2D). The
316 DLS reduced false positives from 520 to 247, while missing only 9 of 162 eyes with ci-DME
317 on OCT. None of the 9 eyes had ci-DME based on retinal thickening. The DLS had an AUC
318 of 0.88 compared to 0.72 for the proof-of-concept DLS.

319

320 Confusion matrices of DLS and expert grades for ci-DME defined by retinal thickening and
321 IRF presence, where available, in all four validation sets is presented in Figure S7.

322 Evaluation in the context of DLS-based DR screening

323 Both original and modified versions of the Krause et al. DLS had similar sensitivities for
324 detecting VTDR ($p < 0.001$ for non-inferiority): 93% vs 92%, respectively. The specificity of
325 the modified DLS was significantly higher than the specificity of the original (69% vs 60%,
326 $p = 0.03$). Thus, the modified Krause et al. DLS had 22% lower false positives for VTDR and
327 8% fewer VTDR referrals, all without loss in sensitivity (Figure 3).

328 Qualitative analysis

329 Finally, we qualitatively analyzed randomly selected instances where the DLS did better or
330 worse compared to retina specialists. Figure 4A shows an eye where HE are present near
331 the fovea whereas the corresponding OCT shows an absence of ci-DME (thickening or
332 fluid). This is a canonical example of false positive reduction achieved by the DLS. Figure 4B
333 illustrates an example where the DLS detects a clear case of ci-DME without any HE on the
334 corresponding CFP. Additional examples of false positive and false negative cases can be
335 found in Figures S5 and S6, respectively. Figure 4C-D demonstrate examples of DLS false
336 negative and false positive, respectively. In both cases the visibility in the macular region
337 appears to be compromised, which could be a potential cause for the DLS errors.

338 Discussion

339

340 We present a DME-DLS that interprets CFP to provide a DME status that is significantly
341 more specific than experts grading CFP for HEs, while retaining non-inferior sensitivity. We
342 also achieved significantly higher performance when compared to a previously reported
343 proof-of-concept version of the DME-DLS. Our results generalized across sites in four
344 countries with diverse populations, to different DME definitions, location of DME, and to
345 subgroups of patients with varying severities of DR, including the mild-to-moderate DR
346 patient population for whom the screening tool will be most applicable. Use of a DLS similar
347 to the one we present here holds particular clinical relevance, as diagnosis within screening
348 centers rely heavily upon CFP, which can be less specific and lead to unnecessary referrals.

349

350 In the UK, OCT surveillance clinics have recently been established as an intermediary stage
351 to refine referrals between screening and specialist centers for patients with mild DR and
352 suspected DME,³⁵ as false positive rates can be as high as 86%.⁶ This is an important cohort
353 to consider as specialty review is generally not required and rescreening at 12–24 months is
354 recommended in the absence of DME.³⁶ While OCT surveillance clinics can be more cost
355 effective relative to direct referral from screening to a specialist center,³⁵ it may not always
356 be logistically feasible, and continues to place significant burden on patients to attend
357 appointments and on expert clinicians to interpret the OCT. Therefore, using the UK
358 validation set, we simulated implementation of the DLS as an ancillary second reader after
359 initial screening has occurred in a cohort of patients that might be referred to the OCT
360 surveillance clinic to identify false positives for DME presence. Our results suggest that the
361 DLS could reduce unwarranted referrals by 80%, while not missing a single positive case.

362

363 In resource-constrained healthcare systems such as in Thailand, the Krause et al. DR
364 grading system has been effective at providing real-time results to patients.³⁷ However,

365 human-centered evaluation observed concerns from both nurses and patients of a false
366 positive result leading to the additional travel burden for referral follow-up, the cost of
367 missing work, and the emotional strain due to a positive result.³⁷ In addition, false positives
368 place significant burden on already overwhelmed secondary care systems. Similarly, the
369 Australian dataset was from an Aboriginal community clinic, where the DR screening occurs
370 with CFP in many geographically isolated communities without specialist services. By
371 reducing false positives for DME, high costs as well as logistical and cultural barriers are
372 avoided when attempting to coordinate follow-up care. For the India validation set, we
373 simulated the application of integrating this highly-specific DME detection DLS for automated
374 DR screening. In this analysis, we were able to reduce false positive VTDR referrals by 22%
375 and overall VTDR referrals by 8% with no statistically significant loss in sensitivity.

376

377 We demonstrate that the DLS can detect DME within 3000 μm of the foveal center and ci-
378 DME affecting the central 500 μm . Anti-vascular endothelial growth factor drugs and steroid
379 implants have demonstrated efficacy in improving visual acuity in patients with ci-DME³⁸ and
380 delaying treatment can lead to suboptimal visual gains.³⁹ Nci-DME is also of significance as
381 it may be a precursor to visually significant ci-DME and should be monitored for risk of
382 progression.⁴⁰ In these patients, focal laser treatment may be indicated to reduce leakage
383 and stabilize visual acuity.^{41,42} Furthermore, our DLS can detect DME defined by thickening
384 or IRF presence. Although these definitions are correlated, we found that only 20–66% of
385 eyes with ci-DME, defined by either thickening or IRF, demonstrated both. Our DLS could be
386 used to triage patients—prioritizing those who meet one or both definitions.

387

388 The performance of the DLS was robust when assessing different degrees of DR severity.
389 DME can be found in eyes at any DR severity level and can run an independent course to
390 DR.⁴³ In patients with mild or moderate DR, there is a risk of missing DME: Wang et al.
391 found over 1 in 4 cases were missed when using CFP labels compared to an OCT reference
392 standard.⁵ Exploratory analysis revealed our DLS to also have a higher sensitivity. Figure 4B

393 and Figure S6 illustrate instances with no hard exudates at the macula but with ci-DME on
394 OCT that was correctly detected by the DLS. Such cases would have been missed if relying
395 upon conventional CFP grading. This could be studied more rigorously with larger and
396 prospectively planned datasets in the future.

397

398 Prior work from our group has shown that the features around the fovea are most relevant
399 for the proof-of-concept version of our DME-DLS.¹⁶ Explainability techniques such as heat
400 maps have been applied, highlighting areas highly correlated with the DLS prediction.⁴⁴

401 Compared to the Krause et al. DLS, which focused on HE locations in the whole 45 degree
402 CFP, the DME-DLS primarily focused on the fovea, leading to superior performance.

403 Another explainability technique in prior work used CycleGAN,⁴⁵ a type of Generative
404 Adversarial Network that transforms negative cases into positive cases and vice versa, to
405 visualize the changes in CFP features that are necessary for the transformation. This
406 approach observed that transformations from DME to no-DME (or vice versa) involved the
407 removal (or addition) of hard exudates and a darkening (or brightening) of the foveal
408 region.⁴⁴ Consistent with these findings, from qualitative analysis, we found that a proportion
409 of incorrect predictions for the current DLS may be attributed to CFP artifacts such as poor
410 contrast or macular shadows that might result from suboptimal pupil dilation (Figure 4C–D).
411 In the future, improved interpretability of the DLS could provide an opportunity for clinicians
412 to learn from and better diagnose DME directly from CFPs.

413

414 In terms of aggregate performance (Australia, India, and Thailand datasets), the DLS met
415 the primary endpoint of superior specificity and non-inferior sensitivity using a pre-specified
416 operating point. However, we note that on a per-dataset level, the ROC curves suggest that
417 the operating points could be further calibrated on a per-site basis to achieve a better trade-
418 off between sensitivity and specificity that accounts for local preferences and resource
419 constraints. Applicable trade-offs are likely to be dependent on local resource constraints
420 and the desired sensitivity. The shape of the ROC curve also varies between the sites, which

421 is likely due to differences in the population distribution. Future work could explore selecting
422 the ideal operating point for a given setting.

423

424 To be clinically applicable and robust, we developed and validated the DLS on images
425 acquired from multiple manufacturers using gender and device-specific thresholds. However,
426 due to site-specific differences in OCT devices and scan protocols, the reference standard
427 for IRF in the secondary analyses differed across datasets in terms of the number of B-
428 scans and area imaged (fovea vs full volumes). Further studies may help understand how
429 these changes in reference standard impact final performance. Furthermore, future work
430 could explore the robustness of the DLS when compared to various thickness thresholds.
431 For example, in the UK, NICE guidelines require point thickness of $\geq 400 \mu\text{m}$ in the central
432 subfield to start treatment. A DLS that can identify different levels of ci-DME could more
433 effectively prioritize patients that would be eligible for treatment.

434 Conclusion

435 This study demonstrates that the DME-DLS can generalize to multiple international
436 populations with an accuracy exceeding both experts and a previous proof-of-concept
437 version of the system. As the prevalence of diabetes is increasing, resulting in more
438 individuals requiring DR screening,⁴⁶ DLS systems are likely to play a significant role in
439 assisting clinicians to ensure timely grading and referrals with both high sensitivity and
440 specificity.⁴⁷ We believe our DLS has most clinical applicability in resource-constrained
441 settings to reduce false positive referrals from screening. Another advantage of our DLS is
442 the ability to triage patients depending on the type of DME, such as ci-DME and nci-DME, as
443 well as thickness and fluid-based definitions. Future work should explore the clinical utility of
444 the DME-DLS through prospective evaluation.

445 **References**

- 446 1. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *The Lancet* 2010;376:124–136.
447 Available at: [http://dx.doi.org/10.1016/s0140-6736\(09\)62124-3](http://dx.doi.org/10.1016/s0140-6736(09)62124-3).
- 448 2. Zhang X, Saaddine JB, Chou C-F, et al. Prevalence of Diabetic Retinopathy in the United
449 States, 2005-2008. *JAMA* 2010;304:649. Available at:
450 <http://dx.doi.org/10.1001/jama.2010.1111>.
- 451 3. Baker CW, Glassman AR, Beaulieu WT, et al. Effect of Initial Management With
452 Aflibercept vs Laser Photocoagulation vs Observation on Vision Loss Among Patients With
453 Diabetic Macular Edema Involving the Center of the Macula and Good Visual Acuity: A
454 Randomized Clinical Trial. *JAMA* 2019;321:1880–1894.
- 455 4. Mackenzie S, Schmermer C, Charnley A, et al. SDOCT imaging to identify macular
456 pathology in patients diagnosed with diabetic maculopathy by a digital photographic retinal
457 screening programme. *PLoS One* 2011;6:e14811.
- 458 5. Wang YT, Tadarati M, Wolfson Y, et al. Comparison of Prevalence of Diabetic Macular
459 Edema Based on Monocular Fundus Photography vs Optical Coherence Tomography.
460 *JAMA Ophthalmology* 2016;134:222. Available at:
461 <http://dx.doi.org/10.1001/jamaophthalmol.2015.5332>.
- 462 6. Wong RL, Tsang CW, Wong DS, et al. Are we making good use of our public resources?
463 The false-positive rate of screening by fundus photography for diabetic macular oedema.
464 *Hong Kong Med J* 2017;23:356–364.
- 465 7. Virgili G, Menchini F, Murro V, et al. Optical coherence tomography (OCT) for detection of
466 macular oedema in patients with diabetic retinopathy. *Cochrane Database Syst Rev*
467 2011:CD008081.
- 468 8. Nguyen QD, Brown DM, Marcus DM, et al. Ranibizumab for diabetic macular edema:

- 469 results from 2 phase III randomized trials: RISE and RIDE. *Ophthalmology* 2012;119:789–
470 801.
- 471 9. Diabetic Retinopathy Clinical Research Network, Wells JA, Glassman AR, et al.
472 Aflibercept, bevacizumab, or ranibizumab for diabetic macular edema. *N Engl J Med*
473 2015;372:1193–1203.
- 474 10. Kim BY, Smith SD, Kaiser PK. Optical coherence tomographic patterns of diabetic
475 macular edema. *Am J Ophthalmol* 2006;142:405–412.
- 476 11. World Health Organization. *World Report on Vision*. World Health Organization; 2019.
477 Available at: <https://www.who.int/publications/i/item/9789241516570>.
- 478 12. Ting DSW, Cheung CY-L, Lim G, et al. Development and Validation of a Deep Learning
479 System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From
480 Multiethnic Populations With Diabetes. *JAMA* 2017;318:2211–2223. Available at: [Accessed
481 April 13, 2021].
- 482 13. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for
483 referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The*
484 *Lancet Digital Health* 2019;1:e35–e44. Available at: [Accessed April 13, 2021].
- 485 14. Nielsen KB, Lautrup ML, Andersen JKH, et al. Deep Learning–Based Algorithms in
486 Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance.
487 *Ophthalmology Retina* 2019;3:294–304. Available at: [Accessed April 13, 2021].
- 488 15. Abràmoff MD, Lou Y, Erginay A, et al. Improved Automated Detection of Diabetic
489 Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Invest*
490 *Ophthalmol Vis Sci* 2016;57:5200–5206.
- 491 16. Varadarajan AV, Bavishi P, Ruamviboonsuk P, et al. Predicting optical coherence
492 tomography-derived diabetic macular edema grades from fundus photographs using deep

- 493 learning. *Nat Commun* 2020;11:130.
- 494 17. Arcadu F, Benmansour F, Maunz A, et al. Deep Learning Predicts OCT Measures of
495 Diabetic Macular Thickening From Color Fundus Photographs. *Invest Ophthalmol Vis Sci*
496 2019;60:852–857. Available at: [Accessed April 13, 2021].
- 497 18. Krause J, Gulshan V, Rahimy E, et al. Grader Variability and the Importance of
498 Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy.
499 *Ophthalmology* 2018;125:1264–1272.
- 500 19. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human
501 graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj*
502 *Digital Medicine* 2019;2. Available at: <http://dx.doi.org/10.1038/s41746-019-0099-8>.
- 503 20. Gulshan V, Rajan RP, Widner K, et al. Performance of a Deep-Learning Algorithm vs
504 Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA Ophthalmol* 2019.
505 Available at: <http://dx.doi.org/10.1001/jamaophthalmol.2019.2004>.
- 506 21. Brown JC, Solomon SD, Bressler SB, et al. Detection of diabetic foveal edema: contact
507 lens biomicroscopy compared with optical coherence tomography. *Arch Ophthalmol*
508 2004;122:330–335.
- 509 22. Sadda SR, Tan O, Walsh AC, et al. Automated detection of clinically significant macular
510 edema by grid scanning optical coherence tomography. *Ophthalmology* 2006;113:1187.e1–
511 12.
- 512 23. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based
513 diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*
514 2018;1:39.
- 515 24. Friedman SM, Almukhtar TH, Baker CW, et al. Topical nepafenec in eyes with
516 noncentral diabetic macular edema. *Retina* 2015;35:944–956.

- 517 25. Wilkinson CP, Ferris FL, Klein RE, et al. Proposed international clinical diabetic
518 retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*
519 2003;110:1677–1682. Available at: [http://dx.doi.org/10.1016/s0161-6420\(03\)00475-5](http://dx.doi.org/10.1016/s0161-6420(03)00475-5).
- 520 26. Public Health England. NHS Diabetic Eye Screening Programme: Overview of patient
521 pathway, grading pathway, surveillance pathways and referral pathways. 2017. Available at:
522 [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachm_ da](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/648658/Diabetic_Eye_Screening_pathway_overviews.pdf)
523 [ta/file/648658/Diabetic_Eye_Screening_pathway_overviews.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/648658/Diabetic_Eye_Screening_pathway_overviews.pdf).
- 524 27. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on
525 Heterogeneous Distributed Systems. *arXiv [csDC]* 2016. Available at:
526 <http://arxiv.org/abs/1603.04467>.
- 527 28. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for
528 diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–1350.
- 529 29. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural
530 Networks. *arXiv [csLG]* 2019. Available at: <http://arxiv.org/abs/1905.11946>.
- 531 30. Fagerland MW, Lydersen S, Laake P. Recommended tests and confidence intervals for
532 paired binomial proportions. *Stat Med* 2014;33:2850–2875.
- 533 31. Liu J-P, Hsueh H-M, Hsieh E, Chen JJ. Tests for equivalence or non-inferiority for paired
534 binary data. *Stat Med* 2002;21:231–245.
- 535 32. Obuchowski NA. On the comparison of correlated proportions for clustered data. *Stat*
536 *Med* 1998;17:1495–1507.
- 537 33. Yang Z, Sun X, Hardin JW. A note on the tests for clustered matched-pair binary data.
538 *Biom J* 2010;52:638–652.
- 539 34. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more
540 correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*

541 1988;44:837–845.

542 35. Leal J, Luengo-Fernandez R, Stratton IM, et al. Cost-effectiveness of digital surveillance
543 clinics with optical coherence tomography versus hospital eye service follow-up for patients
544 with screen-positive maculopathy. *Eye* 2019;33:640–647.

545 36. Public Health England. NHS Diabetic Eye Screening Programme: Grading definitions for
546 referable diseases. 2012. Available at:
547 [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_da](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/582710/Grading_definitions_for_referrable_disease_2017_new_110117.pdf)
548 [ta/file/582710/Grading_definitions_for_referrable_disease_2017_new_110117.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/582710/Grading_definitions_for_referrable_disease_2017_new_110117.pdf).

549 37. Beede E, Baylor E, Hersch F, et al. A Human-Centered Evaluation of a Deep Learning
550 System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: *Proceedings of the*
551 *2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY,
552 USA: Association for Computing Machinery; 2020:1–12. Available at: [Accessed March 31,
553 2021].

554 38. Schmidt-Erfurth U, Garcia-Arumi J, Bandello F, et al. Guidelines for the Management of
555 Diabetic Macular Edema by the European Society of Retina Specialists (EURETINA).
556 *Ophthalmologica* 2017;237:185–222.

557 39. Boyer DS, Nguyen QD, Brown DM, et al. Outcomes with As-Needed Ranibizumab after
558 Initial Monthly Therapy: Long-Term Outcomes of the Phase III RIDE and RISE Trials.
559 *Ophthalmology* 2015;122:2504–13.e1.

560 40. Bressler NM, Diabetic Retinopathy Clinical Research Network, Miller KM, et al.
561 Observational study of subclinical diabetic macular edema. *Eye* 2012;26:833–840. Available
562 at: <http://dx.doi.org/10.1038/eye.2012.53>.

563 41. Scott IU, Danis RP, Bressler SB, et al. Effect of focal/grid photocoagulation on visual
564 acuity and retinal thickening in eyes with non-center-involved diabetic macular edema.
565 *Retina* 2009;29:613–617.

- 566 42. Perente I, Alkin Z, Ozkaya A, et al. Focal laser photocoagulation in non-center involved
567 diabetic macular edema. *Med Hypothesis Discov Innov Ophthalmol* 2014;3:9–16.
- 568 43. Keenan TDL, Johnston RL, Donachie PHJ, et al. United Kingdom National
569 Ophthalmology Database Study: Diabetic Retinopathy; Report 1: prevalence of centre-
570 involving diabetic macular oedema and other grades of maculopathy and retinopathy in
571 hospital eye services. *Eye* 2013;27:1397–1404.
- 572 44. Narayanaswamy A, Venugopalan S, Webster DR, et al. Scientific Discovery by
573 Generating Counterfactuals using Image Translation. *arXiv [csCV]* 2020. Available at:
574 <http://arxiv.org/abs/2007.05500>.
- 575 45. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-
576 Consistent Adversarial Networks. *arXiv [csCV]* 2017. Available at:
577 <http://arxiv.org/abs/1703.10593>.
- 578 46. Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes
579 prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* 2018;138:271–281.
- 580 47. Wong TY, Sabanayagam C. Strategies to Tackle the Global Burden of Diabetic
581 Retinopathy: From Epidemiology to Artificial Intelligence. *Ophthalmologica* 2020;243:9–20.
- 582

583 **Figure Legends**

584 **Figure 1. Study design.** A: The deep learning system (DLS) was trained to take color
585 fundus photography (CFP) as input to predict optical coherence tomography (OCT)-derived
586 diabetic macular edema (DME) presence, using datasets from Thailand, the UK and the US.
587 B: The DLS was evaluated for non-inferiority of sensitivity and superiority of specificity, and
588 compared to experts on datasets from Australia, India, and Thailand. The DLS was
589 separately evaluated on a dataset from the UK. Given the difference in patient population
590 (DR screening in Australia, India and Thailand, vs. pre-screened diabetes patients with a
591 higher likelihood of DME in the UK), these datasets were separately analyzed. Both the DLS
592 and expert graders saw only the CFP. The reference standard was based on measurements
593 from OCT.

594 **Figure 2.** Receiver operating characteristic curves of the DME-DLS, a previously reported
595 proof-of-concept version of the DLS (not generalized), and experts for detecting ci-DME. The
596 threshold for the DLS was pre-specified. The experts graded the presence of hard exudates
597 within 1500 μm . A) Comparison of performance of both DLSs and experts in a combined
598 cohort of the screening datasets from Australia, India and Thailand consisting of eyes with
599 mild or worse DR. ci-DME was defined by central subfield thickness exceeding OCT device
600 specific threshold. B) Comparison of performance of both DLSs in the separate UK
601 validation dataset consisting of patients referred from screening for DME. C) Comparison of
602 performance of both DLSs and experts in the Australia and India datasets on eyes with mild
603 or worse DR. ci-DME was defined by the presence of intraretinal fluid in the central subfield.
604 D) Comparison of performance of both DLSs in the UK validation dataset. Intraretinal fluid
605 presence grades were not available for the Thailand dataset.

606 **Figure 3: Effect of replacing the DME component of the Krause et al. DLS with the**
607 **DME-DLS presented in this paper on the India dataset.** The Krause et al. DLS (left) and
608 the modified Krause et al. DLS (right) have identical rates of severe+DR compared with the

609 adjudicated ground truth (center). Yet, the dark red flow (left) shows the substantially larger
610 number of DME categorizations made by the original Krause et al. DLS compared to the
611 brighter red flow (right) for the modified Krause et al. DLS (i.e. reduced number of false
612 positives by the modified DLS); the sensitivity remains unchanged (as shown by the green
613 bar on the top).

614 **Figure 4. Selected paired CFPs and OCTs for DLS success and failure, as compared with**
615 **retina specialists (RS) grading CFPs.** A) Hard exudates (HE) within 1500 μm , no thickening or
616 fluid on the OCT - detected correctly by the DLS. B) No HE within 1500 μm , thickening and fluid
617 on the OCT - detected correctly by the DLS. C) HE within 1500 μm , thickening and fluid on the
618 OCT - missed by the DLS. D) No HE within 1500 μm , no thickening or fluid on the OCT - false
619 positive reported by the DLS.

Characteristic	Australia	India	Thailand	United Kingdom
Institution	Derbarl Yerrigan Health Service, Perth	Sankara Nethralaya, Chennai	Rajavithi Hospital, Bangkok	Moorfields Eye Hospital, London
Collection dates	July 2013 to October 2020	October 2019 to February 2020	February 2020 to July 2020	August 2014 to September 2018
Population	Diabetic patients presenting for DR screening	Diabetic patients visiting outpatient ophthalmology clinic	Diabetic patients presenting for DR screening	Diabetic patients randomly selected from a cohort referred from the DR screening program for at least one eye with mild DR and maculopathy
Patients	866	168	548	537
Age in years, mean (SD)	54.9 (15.0) n=866	60.0 (8.6) n=168	57.6 (11.2) n=548	55.0 (15.1) n=537
Sex (% female)	52.4% n=454	31.5% n=53	62.2% n=341	41.7% n=224
Eyes (one image per eye)	1692	298	1070	698
Eyes used in analysis after excluding ungradable images, %	90.4% n=1530	98.0% n=292	98.5% n=1054	100.0% n=698
No DR, %	75.6% n=1157	44.2% n=129	87.0% n=917	0.0% n=0
Mild DR, %	3.2% n=49	3.4% n=10	4.0% n=42	100.0% n=698
Moderate DR, %	16.5% n=253	27.7% n=81	6.9% n=73	0.0% n=0
Severe DR, %	1.0% n=15	3.1% n=9	1.3% n=14	0.0% n=0
Proliferative DR, %	3.7% n=56	21.6% n=63	0.8% n=8	0.0% n=0
Central subfield thickness in μm , mean (SD)	238.6 (54.5) n=1530	297.5 (118.8) n=291	270.9 (56.0) n=1054	233.0 (42.5) n=698
ci-DME positive (central subfield thickness \geq threshold), %	3.8% n=58	21.0% n=61	2.5% n=26	5.3% n=37
ci-DME positive (IRF present), %	11.1% n=166	25.0% n=72	Not available	23.8% n=162
ci-DME positive (central retinal thickening and IRF present), %	3.1% n=46	18.4% n=53	Not available	4.8% n=33
DME positive (retinal thickening \geq threshold)	13.1% n=200	39.0% n=114	4.6% n=49	Not available

Table 1. Baseline characteristics of the validation datasets from multiple institutions in Australia, India, Thailand, and the United Kingdom. Abbreviations: ci-DME: center-involving diabetic macular edema; DR: diabetic retinopathy; SD: standard deviation; IRF: intraretinal fluid.

Journal Pre-proof

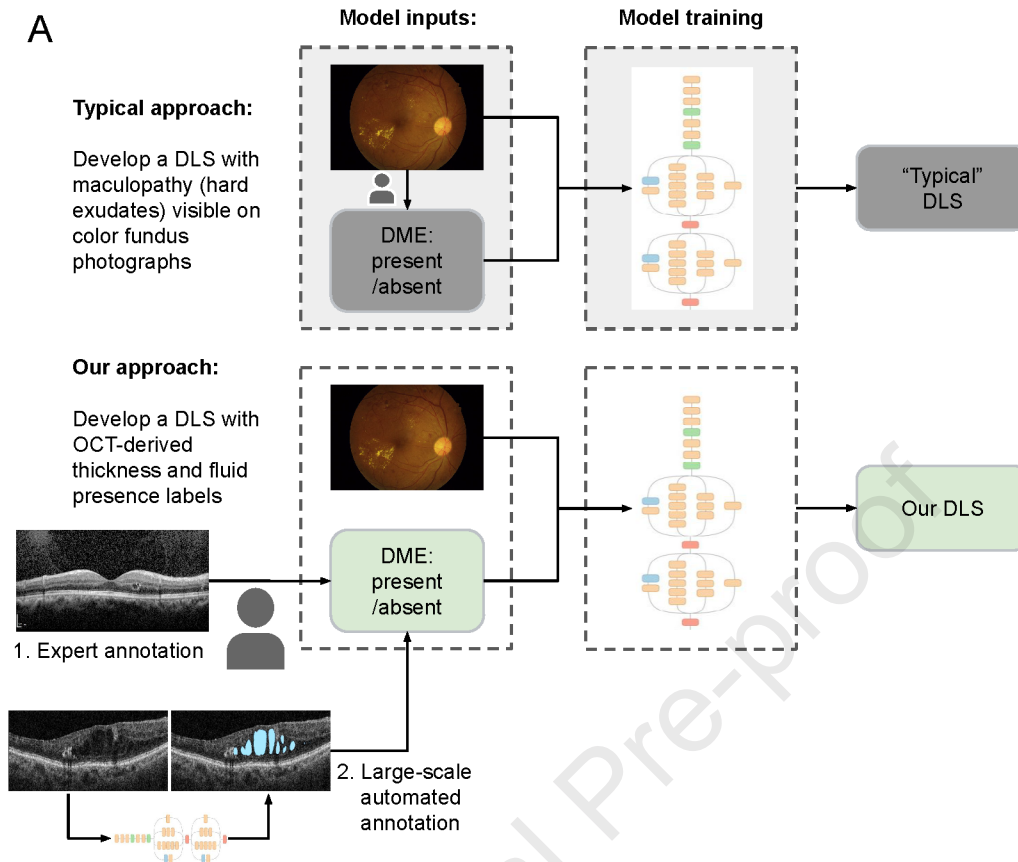
Dataset	Imaging device(s) used	Ground truth for primary analysis (ci-DME based on CST)	Ground truth for secondary analysis	Expert DR & DME grades from CFP
Australia	CFP + OCT: 3D OCT-1 Maestro (Topcon Corp., Tokyo, Japan)	Topcon software	1. Thickness-based ci-DME and nci-DME from Topcon software 2. Majority grade by 3 ophthalmologists for IRF-based ci-DME and DME using the full OCT volume	Single grades by a pool of 7 retina specialists
India	CFP: NFC 700 (Crystalvue, Taoyuan City, Taiwan) or NW400 (Topcon Corp., Tokyo, Japan) OCT: Cirrus HD-OCT 500 (Carl Zeiss Meditec, Dublin, CA)	Zeiss software	1. Thickness-based ci-DME and nci-DME from Zeiss software 2. Adjudicated grade by 1 retina specialist and 1 ophthalmologist for IRF-based ci-DME using the OCT report, with the central B-Scan containing the fovea	3-way adjudicated grades by a pool of 18 experts (13 retina specialists, 2 ophthalmologists, 3 optometrists)
Thailand	CFP: VX-10 (Kowa, Tokyo, Japan) OCT: Spectralis (Heidelberg Engineering, Heidelberg, Germany)	Heidelberg software	1. Thickness-based ci-DME and nci-DME from Heidelberg software 2. IRF grades not available	Single grades by a pool of 5 retina specialists
UK	CFP + OCT: 3D OCT-2000 (Topcon Corp., Tokyo, Japan)	Topcon software	1. Thickness-based ci-DME from Topcon software after manual recentering of ETDRS grid to fovea (nci-DME measurements not available) 2. Majority grade by 3 ophthalmologists for IRF-based ci-DME and DME using the full OCT volume	Screening grade, derived by a 3-level grading system ²⁶ using certified retinal graders in the diabetic eye screening program

Table 2. Sources of ground truths and expert grades for comparison in the validation datasets. CFP: color fundus photograph; ci-DME: center-involving diabetic macular edema; CST: central subfield thickness; DR: diabetic retinopathy; DME: diabetic macular edema; ETDRS, Early Treatment Diabetic Retinopathy Study; IRF, Intraretinal Fluid; OCT: optical coherence tomography.

	Combined	Australia	India	Thailand
Number of images	673	373	163	137
Number of patients	457	247	106	104
Number of images positive for ci-DME	145	58	61	26
Model specificity	80%	86%	77%	66%
Grader specificity	59%	57%	44%	78%
95% CI for difference	[5%, 36%]	[3%, 54%]	[22%, 43%]	[-32%, 7%]
p-value for difference	0.008	0.030	<0.001	0.201
Model sensitivity	81%	71%	84%	100%
Grader sensitivity	70%	66%	72%	77%
p-value for non-inferiority	<0.001	0.007	0.002	0.013
95% CI for difference	[2%, 20%]	[-7%, 17%]	[-4%, 26%]	[-6%, 52%]
p-value for difference	0.014	0.402	0.127	0.121

Table 3. DLS sensitivity and specificity compared to expert grades for detecting ci-DME in eyes with at least mild DR. The experts graded for the presence of hard exudates within 1500 μm . ci-DME was defined by central subfield thickness exceeding OCT device specific threshold. Abbreviations: ci-DME, center-involving Diabetic Macular Edema; CI, Confidence Interval

A



B

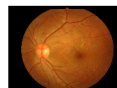
DLS development datasets: 🇹🇭 Thailand, 🇬🇧 UK and 🇺🇸 US

DLS validation datasets: 🇦🇺 Australia, 🇮🇳 India, 🇹🇭 Thailand and 🇬🇧 UK

Comparison: Human experts grading for hard exudates within 1500 microns on color fundus photographs

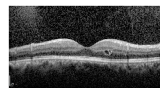
Image for DLS input and expert review:

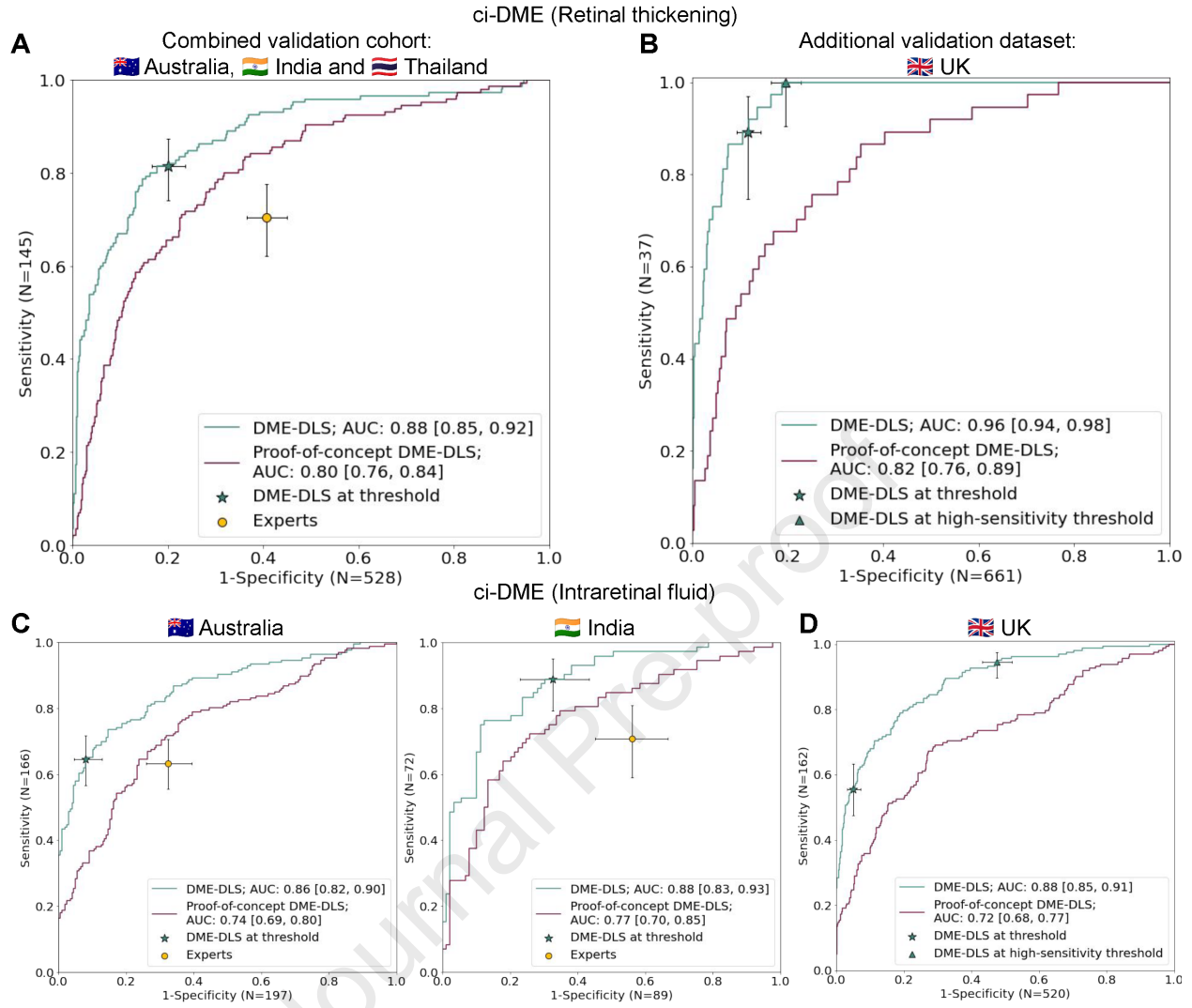
Color fundus photograph

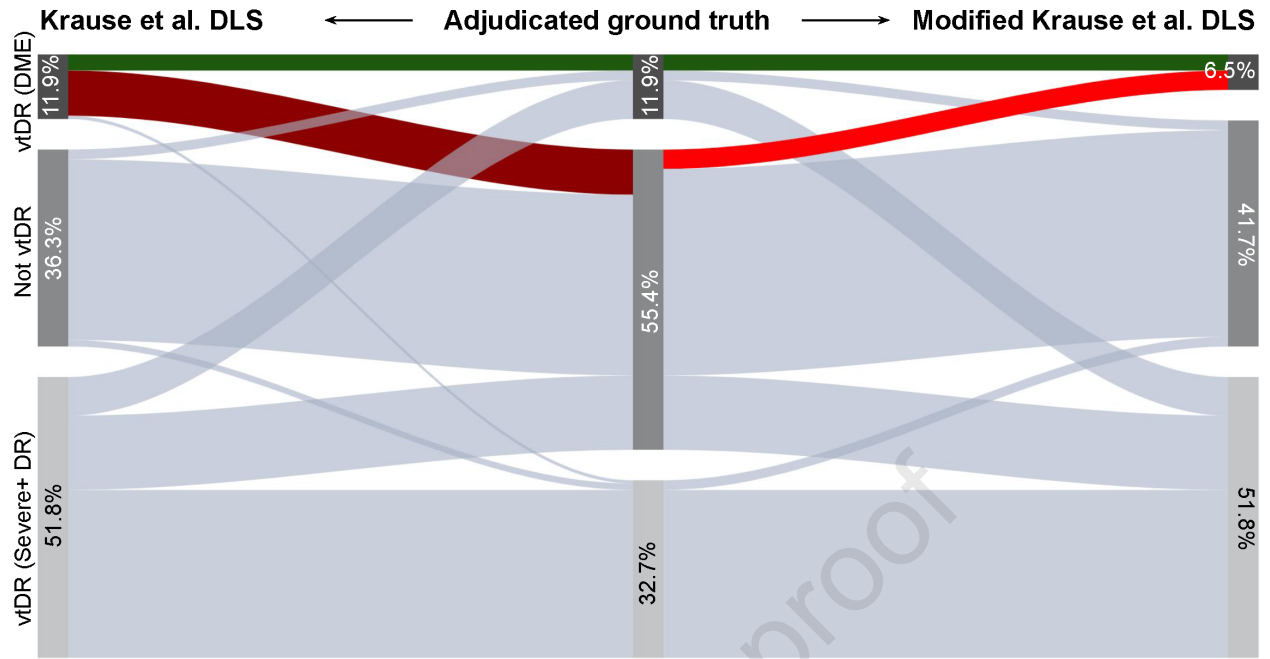


Reference standard:

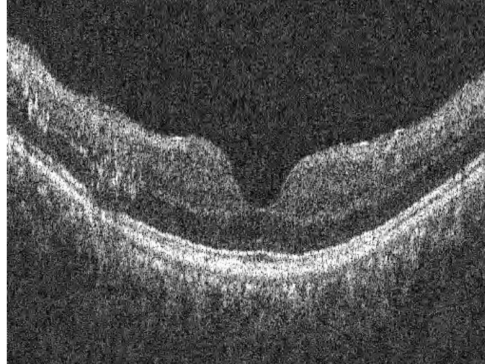
CST \geq 300 microns from optical coherence tomography







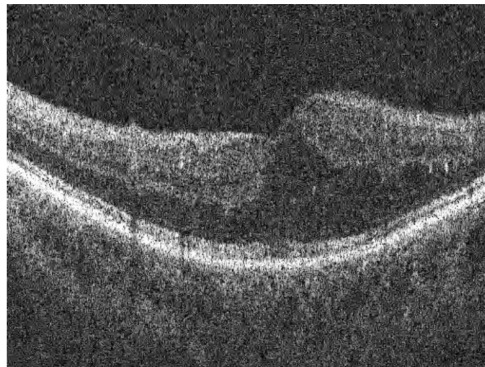
A



Age: 44
Sex: Male
CST: 179 μm
IRF: No
ci-DME: No

RS: Positive for
maculopathy / HE
DLS: Negative for ci-DME

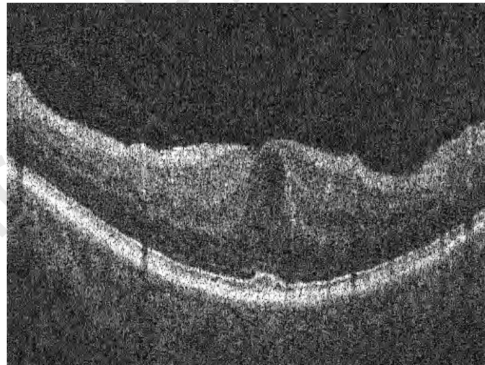
B



Age: 83
Sex: Female
CST: 353 μm
IRF: Yes
ci-DME: Yes

RS: Negative for
maculopathy / HE
DLS: Positive for ci-DME

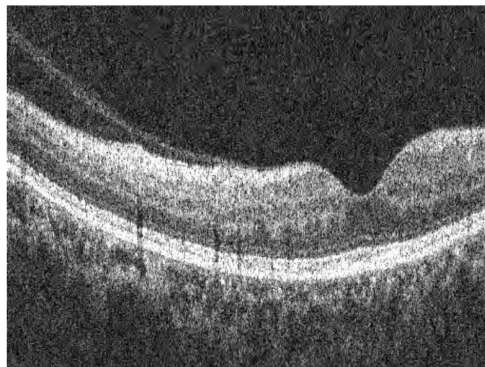
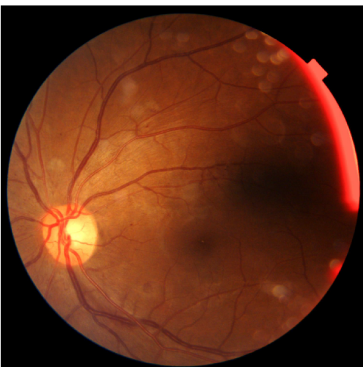
C



Age: 63
Sex: Male
CST: 361 μm
IRF: Yes
ci-DME: Yes

RS: Positive for
maculopathy / HE
DLS: Negative for ci-DME

D



Age: 41
Sex: Female
CST: 267 μm
IRF: No
ci-DME: No

RS: Negative for
maculopathy / HE
DLS: Positive for ci-DME

Precis

A deep learning system was trained to predict OCT-derived diabetic macular edema grades from color fundus photographs, and evaluated on international datasets. It achieved a superior specificity and comparable sensitivity to experts grading photographs.

Journal Pre-proof