Automated detection of plus disease in retinopathy of prematurity using deep learning: A retrospective cohort study

SK Wagner[1,2,3], B Liefers[1], M Radia[3], G Zhang[1], R Struyven[1,2,3], L Faes[1,3], J Than[3], S Balal[3], C Hennings[3], C Kilduff[3], P Pooprasert[3], S Glinton[1], M Arunakirinathan[3], P Giannakis[4], IZ Braimah[5], ISH Ahmed[6,7], M Al-Feky[8,9], H Khalid[3,10], D Ferraz[2,11], J Vieira[12], R Jorge[12], S Husain[13,14], J Ravelo[13], A Hinds[3], R Henderson[15,16], HI Patel[3,17], S Ostmo[18], JP Campbell[18], N Pontikos[1,2,3], PJ Patel[1,2,3], PA Keane[1,2,3], G Adams[1,3], K Balaskas[1,2,3]

[1]NIHR Moorfields Biomedical Research Centre, London, UK, [2]Institute of Ophthalmology, UCL, London, UK, [3]Moorfields Eye Hospital NHS Foundation Trust, London, UK, [4]Institute of Health Sciences Education, Queen Mary University of London, London, UK, [5]Lions International Eye Centre, Korle-Bu Teaching Hospital, Accra, Ghana, [6]Faculty of Medicine, Alexandria University, Alexandria, Egypt, [7]Alexandria University Hospital, Alexandria, Egypt, [8]Department of Ophthalmology, Ain Shams University Hospitals, Cairo, Egypt, [9]Watany Eye Hospital, Cairo, Egypt, [10]Tanta University, Tanta, Egypt, [11]D'Or Institute for Research and Education, Sao Paulo, Brazil, [12]Department of Ophthalmology, RIbeirão Preto Medical School, University of São Paulo, RIbeirão Preto, Brazil, [13]Neonatology Department, Homerton University Hospital NHS Foundation Trust, London, UK, [14]The Blizard Institute, Queen Mary University of London, London, UK, [15]Clinical and Academic Department of Ophthalmology, Great Ormond Street Hospital for Children, London, UK, [16]UCL-GOSH Institute of Child Health, UCL, London, UK, [17]The Royal London Hospital, Barts Health NHS Trust, London, UK, [18]Department of Ophthalmology, Oregon Health & Science University, Portland, OR, USA

Purpose

Retinopathy of prematurity (ROP), a leading cause of childhood blindness, is typically diagnosed through interval screening by paediatric ophthalmologists. However, improved survival of premature neonates coupled with a scarcity of available experts has raised concerns about the sustainability of such approaches. We developed and validated bespoke and code-free deep learning-based (CFDL) classifiers for plus disease, a hallmark of ROP, in an ethnically diverse population in London, United Kingdom. CFDL is not reliant on the availability of expertly trained data scientists and local high performance-computing, thus being of particular potential benefit for low resource health-care settings.

Methods

Consecutive infant retinal images were acquired between January 1st 2008 and January 31st 2018, as part of routine care, at Homerton University Hospital NHS Foundation Trust, London, UK. Images were acquired on all babies within the designated study period who fit the 2008 RCOphth screening criteria for ROP (note the RCOphth guidelines were changed in March 2022).

Reference standard

Each image was independently graded by two junior ophthalmologists with three years of ophthalmology experience as either ungradable, normal, pre-plus or plus disease. Disagreements were resolved by a senior paediatric ophthalmologist, who was masked to the grading of the junior ophthalmologist.

Model development

We exercised two main approaches for model development – traditional bespoke and CFDL.

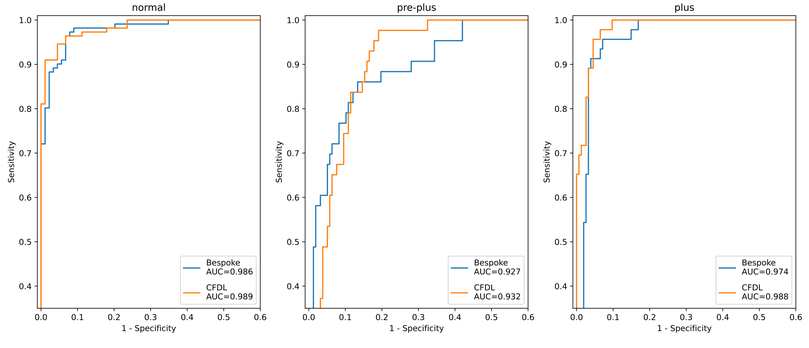
Model validation

The internal test set consisted of 200 images from 200 eyes of 112 infants. The reference standard was the majority class of three senior paediatric ophthalmologist consultants (range of experience: 21-45 years). In the case of equal disagreement between the three classes, we took the most severe label.

External validation was on four international datasets (Table). Three datasets were acquired using the same imaging device as per the development dataset (Retcam) and one using the 3nethra neo device (Forus Health Incorporated, Bengaluru, India).

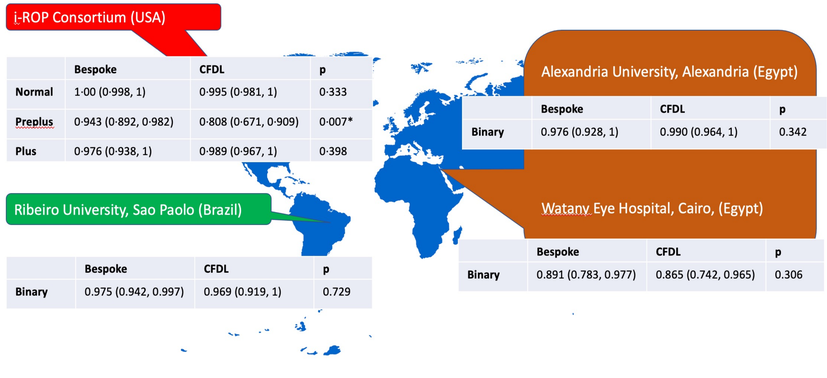
	Internal		External			
	Development	Test	i-ROP	Brazil[1]	Egypt Retcam	Egypt 3nethra
Normal	5771 (94·0%)	111 (55.5%)	54 (54%)	72 (78.3%)	13 (28.9%)	30 (29.7%)
Pre-plus	235 (3·8%)	43 (21·5%)	31 (31%)	20 (21.7%)	32 (71.1%)	71 (70.3%)
Plus	135 (2·2%)	46 (23·0%)	15 (15%)			
Total	6141	200	100	92	45	101

Results



The bespoke model achieved AUCs of 0·986 (0·973-0·996), 0·927 (0·884-0·962) and 0·974 (0·951-0·991) for the discrimination of normal, pre-plus and plus disease versus others respectively.

For the task of normal, pre-plus and plus disease detection versus others, the AUCs of the CFDL model on the internal test set were 0·989 (0·979, 0·997), 0·932 (0·896, 0·964) and 0·988 (0·976, 0·996) respectively



	Bespoke	CFDL	p
Normal	1·00 (0·998, 1)	0·995 (0·981, 1)	0·333
Preplus	0·943 (0·892, 0·982)	0·808 (0·671, 0·909)	0·007*
Plus	0·976 (0·938, 1)	0·989 (0·967, 1)	0·398

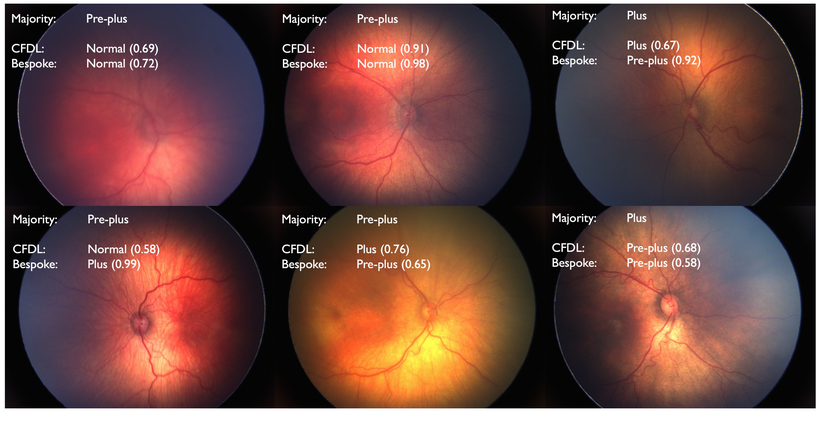
	Bespoke	CFDL	p
Binary	0.975 (0.942, 0.997)	0.969 (0.919, 1)	0.729

	Bespoke	CFDL	p
Binary	0.976 (0.928, 1)	0.990 (0.964, 1)	0.342

	Bespoke	CFDL	p
Binary	0.891 (0.783, 0.977)	0.865 (0.742, 0.965)	0.306

Both models generalised well to external validation test sets acquired using the Retcam for discriminating normal from abnormal (bespoke range: 0.975-1.00, CFDL range: 0.969-0.995). The CFDL model was inferior to the bespoke model on discriminating pre-plus disease from others in the US dataset (bespoke 0·943, 95% CI: 0·892, 0·982, CFDL: 0·808, 95% CI: 0·671, 0·909, p=0·007). Performance also reduced when tested on the 3nethra imaging device (bespoke: 0·891, 95% CI: 0·783, 0·977, CFDL: 0·865, 95% CI: 0·742, 0·965).

Results



Most misclassifications were where the majority considered the image as pre-plus disease. Example misclassifications in the internal test set are shown above.

Discussion

Performance of both bespoke and CFDL models for the discrimination of normal versus plus/pre-plus disease was high. Both models generalised well on international external validation datasets using the same imaging device (Retcam) however the CFDL model performed less well than the bespoke on the task of detecting pre-plus disease. Both models experienced a performance drop when tested on a separate imaging device, the 3nethra neo. While deployment of such models requires a thorough evaluation of effectiveness, our diagnostic accuracy results highlight the potential for automated plus disease diagnosis in UK neonatal units currently leveraging Retcam-based neonatal screening approaches as well as use of CFDL platforms for in-house development of such tools.

We are now investigating wider UK-based prospective validation of this model and undertaking human-computer interaction experiments to assess whether such models can

Acknowledgements

The authors thank the patients, whose data were used as part of this study, as well as their families. This research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology.

Correspondence: s.wagner@ucl.ac.uk