tvst

Artificial Intelligence

Multicenter Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Plus Disease in ROP

Amelia Bai[1–3], Shuan Dai[1,3,4], Jacky Hung[2], Aditi Kirpalani[5], Heather Russell[5,6], James Elder[7,8], Shaheen Shah[9], Christopher Carty[10,11], and Zachary Tan[12]

[1] Department of Ophthalmology, Queensland Children's Hospital, South Brisbane, Queensland, Australia
[2] Centre for Children's Health Research, South Brisbane, Queensland, Australia
[3] School of Medical Science, Griffith University, Southport, Queensland, Australia
[4] University of Queensland, St Lucia, Queensland, Australia
[5] Department of Ophthalmology, Gold Coast University Hospital, Southport, Queensland, Australia
[6] Bond University, Robina, Queensland, Australia
[7] Department of Ophthalmology, Royal Women's Hospital, Parkville, Victoria, Australia
[8] University of Melbourne, Parkville, Victoria, Australia
[9] Mater Misericordiae, South Brisbane, Queensland, Australia
[10] Griffith Centre of Biomedical and Rehabilitation Engineering (GCORE), Menzies Health Institute Queensland, Griffith University, Southport, Australia
[11] Department of Orthopaedics, Children's Health Queensland Hospital and Health Service, Queensland Children's Hospital, South Brisbane, Australia
[12] Aegis Ventures, Sydney, New South Wales, Australia

Correspondence: Shuan Dai, Director of Ophthalmology, Children's Health Queensland Hospital and Health Service, Queensland Children's Hospital, Level 7d, Surgical Directorate, 501 Stanley St, South Brisbane, 4101 QLD, Australia. e-mail: shuan.dai@health.qld.gov.au

Received: February 9, 2023
Accepted: June 30, 2023
Published: August 14, 2023

Keywords: artificial intelligence; retinopathy of prematurity; deep learning; diagnostic screening programs

Citation: Bai A, Dai S, Hung J, Kirpalani A, Russell H, Elder J, Shah S, Carty C, Tan Z. Multicenter validation of deep learning algorithm ROP.AI for the automated diagnosis of plus disease in ROP. Transl Vis Sci Technol. 2023;12(8):13, https://doi.org/10.1167/tvst.12.8.13

Purpose: Retinopathy of prematurity (ROP) is a sight-threatening vasoproliferative retinal disease affecting premature infants. The detection of plus disease, a severe form of ROP requiring treatment, remains challenging owing to subjectivity, frequency, and time intensity of retinal examinations. Recent artificial intelligence (AI) algorithms developed to detect plus disease aims to alleviate these challenges; however, they have not been tested against a diverse neonatal population. Our study aims to validate ROP.AI, an AI algorithm developed from a single cohort, against a multicenter Australian cohort to determine its performance in detecting plus disease.

Methods: Retinal images captured during routine ROP screening from May 2021 to February 2022 across five major tertiary centers throughout Australia were collected and uploaded to ROP.AI. AI diagnostic output was compared with one of five ROP experts. Sensitivity, specificity, negative predictive value, and area under the receiver operator curve were determined.

Results: We collected 8052 images. The area under the receiver operator curve for the diagnosis of plus disease was 0.75. ROP.AI achieved 84% sensitivity, 43% specificity, and 96% negative predictive value for the detection of plus disease after operating point optimization.

Conclusions: ROP.AI was able to detect plus disease in an external, multicenter cohort despite being trained from a single center. Algorithm performance was demonstrated without preprocessing or augmentation, simulating real-world clinical applicability. Further training may improve generalizability for clinical implementation.

Translational Relevance: These results demonstrate ROP.AI's potential as a screening tool for the detection of plus disease in future clinical practice and provides a solution to overcome current diagnostic challenges.

Introduction

Retinopathy of prematurity (ROP) is a sight-threatening vasoproliferative retinal disease affecting premature infants. Those born weighing less than 1250 g or at less than 31 weeks of gestation are most at risk for developing severe ROP, which, if left untreated, can cause retinal detachment and permanent blindness.[1] A crucial feature in treatment-requiring ROP is the presence of plus disease, defined as dilation and tortuosity of retinal vessels in the posterior retina.[2] Landmark studies have established that severe ROP through the early detection of plus disease can be effectively treated with cryotherapy, laser photocoagulation, or intravitreal injections of anti-vascular endothelial growth factor.[1,3,4] It is, therefore, essential that the screening and diagnosis of plus disease be conducted accurately and efficiently to provide timely treatment to prevent the severe sequelae of this treatable disease.

Multiple challenges exist in the timely screening and diagnosis of ROP. Most notably, the diagnosis of plus disease is invariably a subjective diagnosis dependent on the clinician's decision at the time of screening. Despite clear international guidelines for the classification of ROP,[2] significant intraclinician and interclinician variability remain.[5] Moreover, significant geographic variation in plus disease diagnosis has also been reported, emphasizing the inconsistency in ROP diagnosis.[6] Second, ROP screening is labor intensive and requires extensive training and experience before an ophthalmologist can become proficient at diagnosis. Infants undergoing ROP screening also require repeated examinations, often weekly, to avoid missing treatable disease, all of which amounts to a considerable time burden. Improvements in neonatal care contribute to a major challenge in timely screening as the rates of preterm birth survival increase.[7] As a result, the demand for ROP screening has increased to an extent where access to expert ophthalmologists experienced and capable of examination and diagnosis is limited.[8] Finally, the shortage of ophthalmologists in regional and rural centers means higher health care expenses to provide screening for infants from these areas. The high cost of transporting premature infants to tertiary care centers adds to the logistical difficulties of screening and contributes to the financial burden on health care centers and affected families.[9]

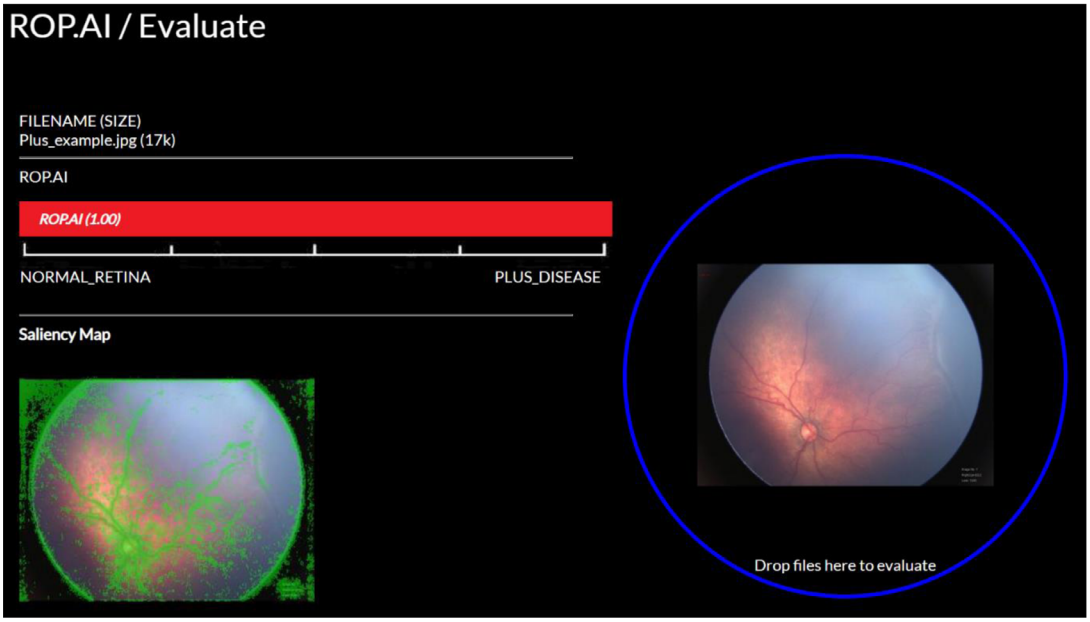
These challenges have incentivized research into large-scale automated screening systems to provide quantitative and objective diagnoses for ROP. The use of artificial intelligence (AI) deep learning technologies has gained particular popularity owing to its ability to self-learn from training image sets to perform certain tasks.[10] Through the use of convolutional neural networks, deep learning algorithms can process multiple features of images to create an architecture of feature maps that together formulate a particular pattern of recognition. As more labelled images are fed into the algorithm, the error between the algorithm diagnosis and the labelled image diagnosis is computed by the AI, and the algorithm is refined to minimize the error. Thus, the larger the number of images used to train the machine, the smaller the error of its diagnostic output. Because AI is self-taught, there lies a major black box issue, where image features recognized by an algorithm are unknown to the user.[10] For this reason, there is hesitancy among clinicians to entrust screening and diagnosis to an AI algorithm. Therefore, AI studies intended to be used for ROP screening must have robust study designs with large datasets to train algorithms. Additionally, external validation with an image set new to the training set is crucial to validate precisely the performance of an algorithm to determine the generalizability into clinical practice.

Groups in the Unites States[11,12] and China[13,14] have developed automated AI systems for the diagnosis of plus disease in ROP; however, no systems have been developed using an Australian cohort. Additionally, limited studies have been validated externally against a geographically novel dataset to ensure reproducible results in a separate population. The deep learning algorithm, ROP.AI, developed from retinal images collected from a single center in New Zealand, has been able to achieve high sensitivity, specificity, and accuracy in detecting plus disease compared with an expert ophthalmologist (96.6%, 98%, and 97.3%, respectively).[15] This system, however, has been trained off one expert ophthalmologist's grading of retinal images. Given the subjectivity of ROP diagnosis, a single expert as the reference standard may not be representative of the current diagnostic standard. The generalizability of ROP.AI's performance to a new population is, therefore, unknown. The aim of this study is to externally validate the performance of ROP.AI against retinal images collected from five different centers across Australia and compare their diagnostic performance in detecting plus disease with five expert ophthalmologists as human graders.

Methods

National ethics approval from the Children's Health Queensland Hospital and Health Service Human Research Ethics Committee was obtained

A:



B:

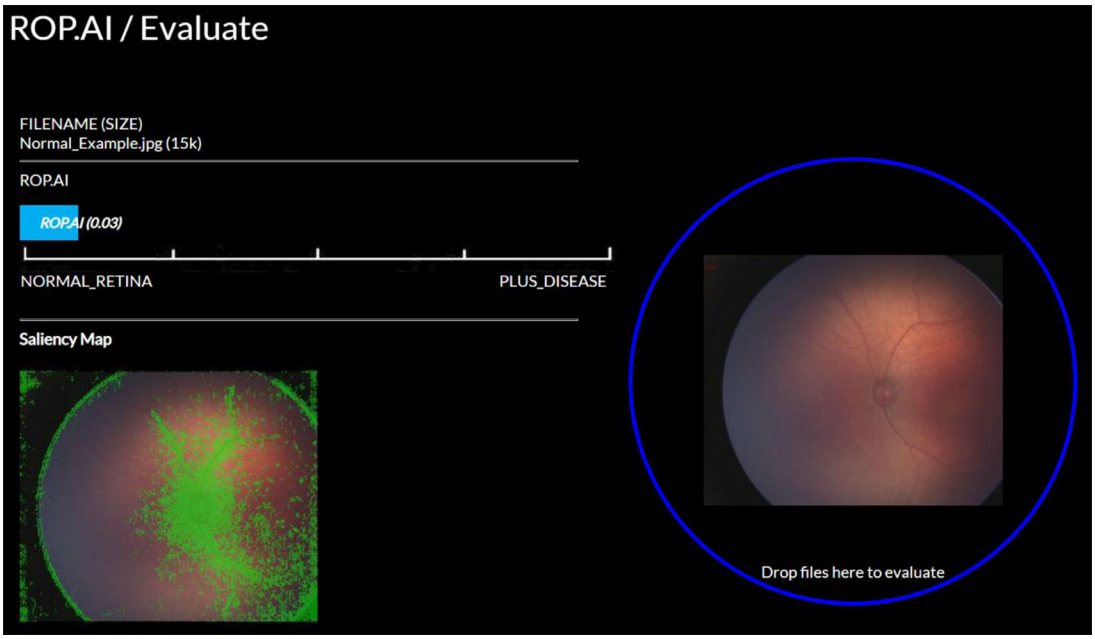


Figure 1. The ROP.AI platform used to upload retinal images for AI analysis. Corresponding AI evaluation of the retinal image is seen on the bar on the left. An output between 0 and 1 is provided by the AI with a saliency map for each image uploaded seen bottom left. (**A**) Example analysis of a plus disease image graded by ROP.AI as 1.00 (plus disease). (**B**) Example analysis of a normal image graded by ROP.AI as 0.03 (normal).

prior to the commencement of this study (HREC/20/QCHQ/62358). All research was conducted in accordance with the Declaration of Helsinki of 1975. A waiver for patient consent was granted by the same ethics committee. A total of 8052 retinal images taken with RetCam 3 (Natus Medical Incorporated, Middletown, WI) were retrospectively collected and deidentified across five different centers in Australia from May 2021 to February 2022 to form an external validation set. The large sample size was chosen to exceed the reported average of 600 external validation images in previous studies.[16] Retinal images were

captured as per the standard ROP imaging protocol to include the posterior retina, nasal retina, and temporal retina and were captured between January 2018 and February 2022. Sites included the Royal Brisbane and Women's Hospital, Queensland; Mater Misericordiae, Queensland; Queensland Children's Hospital, Queensland; Gold Coast University Hospital, Queensland; and The Royal Women's Hospital, Victoria. Images supplied by the five centers were deemed gradable, as indicated by their gradeability in clinical practice, by the individual clinician who provided the images. All images that included the optic nerve head anywhere in the captured Retcam image were accepted. Images were not edited or adjusted and did not undergo augmentation. This protocol differed from the development data, which required images to be centered around the optic disc, high image clarity through strict inclusion criteria, and image augmentation as published previously. Images were chosen at random by one of the five experts and were all collected during routine ROP screening as per each hospital network's ROP guidelines (Appendix 1). The clinician diagnosis for each image, corresponding with their actual clinical diagnosis during routine ROP screening, was collected and images were labelled as either normal, pre-plus, or plus disease. Clinical diagnoses were determined off Retcam images only and patient demographic data were available to the expert at the time of grading. Images were supplied by one of five experts, all of whom are practicing pediatric ophthalmologists with more than 50 years of combined experience in ROP diagnosis.

The performance of the ROP.AI algorithm to detect plus disease and plus and pre-plus combined, through the cloud-based platform MedicMind (https://ai.medicmind.tech), was evaluated against all 8052 images (Fig. 1). All images were naïve to the algorithm and had not been used previously for algorithm training. All images were uploaded to the ROP.AI platform by J.H. Using MedicMind's TensorFlow's Inception-v3 convolutional neural network and RMSProp optimizer (weight decay factor 0.00004, momentum 0.9), ROP.AI determined the probability of diagnosis for each image. Violation of the independence assumption was not accounted for during image analysis. The subsequent diagnosis was recorded and compared against its corresponding clinician diagnosis. Statistical performance for the classifier was measured by calculating sensitivity, specificity, negative predictive value, and area under the receiver operating characteristic curve. ROP.AI performance in diagnosing plus disease vs no plus disease and plus disease and pre-plus combined vs normal was compared with the reference standard diagnosis.

ROP.AI produced a probability value between 0 and 1 after evaluation of retinal images. A default threshold of greater than 0.5 was used initially to determine an image with plus disease. Given the prospect ROP.AI holds to be used as a screening algorithm, the operating point of 0.5 was further optimized to produce high sensitivity and negative predictive value to decrease the likelihood of missed diagnoses. The algorithm was retested against the normal and plus disease external validation set at 0.01 operating point intervals between 0.3 and 0.7. All five ROP experts agreed that an operating point of 0.38 would be used because it yielded an optimal sensitivity and negative predictive value without a total compromise on specificity. It was also chosen to produce comparative results in this external test set compared with the original performance during algorithm development as published previously.[15] The sensitivity, specificity, and negative predictive value was re-calculated for this cut-off point. A random selection of 90 misclassified images were reviewed to identify any causes for algorithm misinterpretation.

Results

A total of 8052 images from 925 individual eyes were analyzed, of which 5879 images (73%) were normal,

Table. Distribution of Images per Center Including the Total Number of Images Supplied, The Average Number of Images per Eye, Number of Plus Disease, Pre-Plus Disease, and Normal Images

Site	Total Images	Avg Image/Eye	Plus Disease	Pre-Plus	Normal
GCUH	4241	10.29	135	782	3324
MM	1785	11.52	521	405	859
QCH	167	6.42	7	66	94
RBWH	1522	5.01	54	0	1468
RWH	337	12.04	68	135	134

GCUH, Gold Coast University Hospital; MM, mater misericordiae; QCH, Queensland Children's Hospital; RBWH, Royal Brisbane and Women's Hospital; RWH, Royal Women's Hospital.

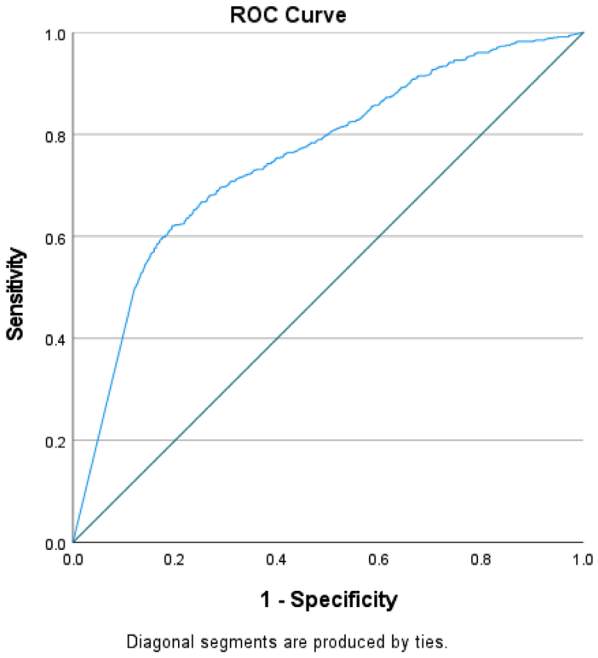


Figure 2. Receiver operating characteristic curve (ROC) curve for ROP.AI diagnosis of plus disease.

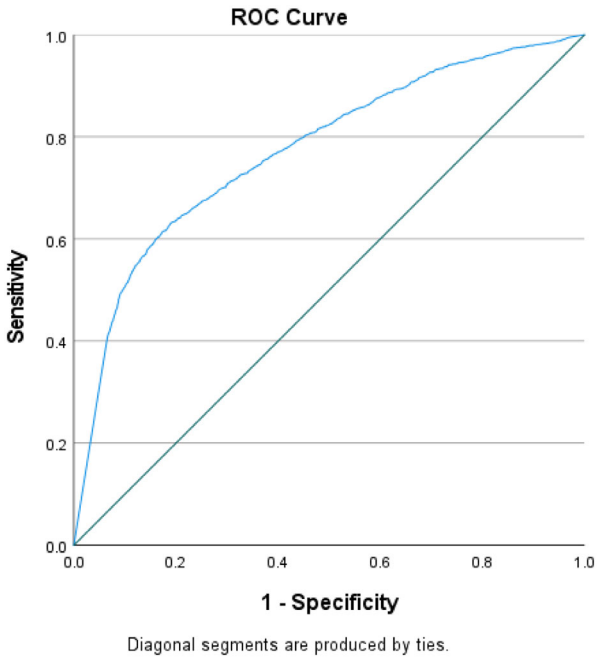


Figure 3. Receiver operating characteristic curve (ROC) curve for ROP.AI diagnosis of pre-plus and plus disease combined.

1388 images (17.2%) had pre-plus disease, and 785 images (9.7%) had plus disease. Each center provided on average 1610 images (range, 167–4241 images). The number of images per center and per diagnosis are listed in Table. The average gestational age of infants at birth was 27.74 ± 2.82 weeks and the average birth weight was 1054.76 ± 378.90 g. The average age of infants at the time of screening was 9.32 ± 4.90 weeks after birth. The New Zealand trained ROP.AI algorithm produced an area under the receiver operating characteristic curve of 0.75 for the detection of plus disease (Fig. 2) and 0.77 for detection of pre-plus and plus disease combined (Fig. 3).

Sensitivity, specificity, and negative predictive value at the algorithm default cut-off point of 0.5 were 78% (95% confidence interval [CI], 75–81), 54% (95% CI, 53–55), and 96% (95% CI, 95–96), respectively, for detecting plus disease. The sensitivity, specificity, and

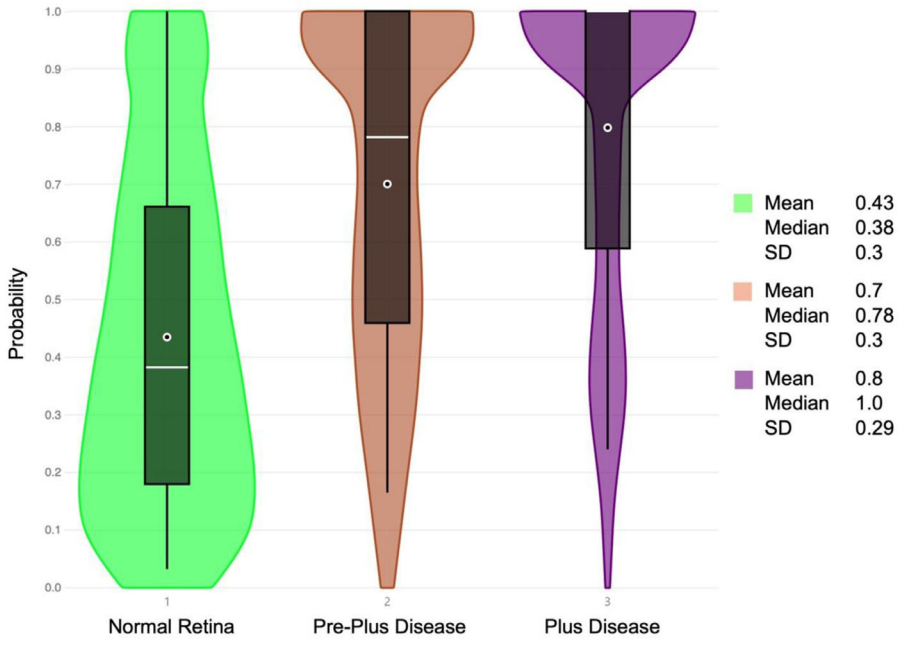


Figure 4. Violin plot for the distribution of algorithm performance in normal, pre-plus, and plus disease fundal images.

negative predictive value in detecting pre-plus and plus disease combined were 76% (95% CI, 75–78), 61% (95% CI, 60–63), and 88% (95% CI, 87–88), respectively. The average outputs produced by the algorithm for normal, pre-plus, and plus disease images were 0.43 ± 0.3, 0.7 ± 0.3, and 0.8 ± 0.29, respectively. The distributions of these probability outputs are illustrated in Figure 4.

After operating point optimization to 0.38, the sensitivity, specificity, and negative predictive value improved to 84% (95% CI, 81–86), 43% (95% CI, 42–44), and 96% (95% CI, 95–97) for the detection of plus disease and 83% (95% CI, 81–85), 49% (95% CI, 48–50), and 89% (95% CI, 88–90) respectively for the detection of pre-plus and plus disease combined. Images that were misclassified seemed to be of darker fundus or slightly blurred (Appendix 2); however, no objective cause was measured.

Discussion

This study evaluates the performance of the New Zealand trained ROP.AI algorithm for diagnosing ROP plus disease and pre-plus disease on a novel set of retinal images from five centers across Australia. We found that the original ROP.AI algorithm performs comparatively well against these new images, and further training may support its potential for application into real-world clinical use. The key findings are that (1) ROP.AI is able to diagnose plus disease on a geographically novel test set, (2) despite being trained to detect plus disease only, results support ROP.AI's ability to distinguish pre-plus disease, and (3) ROP.AI produced diagnostic ability to determine plus disease and plus and pre-plus combined without preprocessing or augmentation, simulating realistic clinical practice for future applicability.

The international classification system for diagnosing ROP provides guidelines on the diagnosis of ROP through the identification of key features on retinal exam.[2] Despite this, inconsistency in ROP and plus disease diagnosis remains prominent,[5] and statistically significant geographic variation in plus disease diagnosis has also been reported.[6] For this reason, an objective screening system such as ROP.AI holds potential to decreased clinically significant management differences and improve outcomes for premature infants. However, it is crucial that the algorithm can overcome both geographic variability and interexpert variability. Our study is unique in that it provides a large test set of more than 8000 images completely naive to the ROP.AI training set with both grader variability (four new expert human graders not previously used for ROP.AI training) and geographic variability (five new centers across Australia). To the best of our knowledge, this automated deep learning algorithm is the first to have undergone large-scale external validation without image preprocessing or augmentation in ROP. Our results reveal ROP.AI performance with a sensitivity and negative predictive value of 84% and 96%, respectively, in the diagnosis of plus disease at the prespecified cut-off point 0.38. Despite being trained from a single center with only one clinician as the reference standard, ROP.AI's performance on this external test set has been comparable with those reported by other groups with a much smaller number of external images.[17–20] These findings support ROP.AI's potential of reproducibility into the clinical setting with potential diagnostic output despite analyzing images it has never been exposed to before.

Although ROP.AI was only trained to recognize plus disease, the algorithm has been able to detect plus and pre-plus combined with a sensitivity and negative predictive value of 83% and 89%. This finding supports the notion that ROP severity remains on a spectrum of retinal vascular changes.[2] The detection of pre-plus disease is clinically important because it warrants close monitoring in preparation for treatment requiring disease. The ability for ROP.AI to detect plus and pre-plus combined was lower when compared with its performance in detecting plus disease; however, this result may be expected, because it was only trained for the detection of plus disease. Future studies, however, should implement algorithm training for the detection of pre-plus disease to further distinguish abnormal retinal vascular patterns from normal retinal images. This metric will have positive clinical implications when used as a screening tool to further categorize disease risk and follow-up time periods.

Another unique feature of the current study is that the images used to validate ROP.AI were not preselected, augmented, or preprocessed before being tested. Unlike other AI studies, we did not limit images to only the posterior pole and accepted any field of view if the optic disc was visible. These techniques are unique to our study because it is well-recognized among AI that high-quality images correlate with high-quality diagnoses and smaller algorithm errors.[10] Meticulous exclusion of poor-quality images and restricted inclusion criteria, however, may limit the applicability of AI algorithms in the real-world clinical setting. It is for this reason that our study accepted a quality and scope of images corresponding with those taken in the clinical setting so that the validation of our algorithm performance may equate with its real-life performance. These factors may have contributed to the low specificities we obtained (43% for the diagnosis of plus disease, 49% for plus and pre-plus disease combined), as well

as the decreased overall performance when compared with the initial development of ROP.AI.

In a disease such as ROP, which holds devastating sight-threatening consequences if treatment is delayed, high sensitivity and negative predictive values are most crucial to avoid missed diagnoses. As a result, low specificity may occur as a compromise, as it does here after optimization for sensitivity and negative predictive value. In practicality, the implementation of ROP.AI as a screening tool to triage at-risk patients will still hold potential to decrease health care costs, despite low specificities. Further algorithm training, however, may overcome this limitation and should be considered before implementation into clinical practice.

This study has several limitations to consider. First, images were collected as part of routine clinical screening by five human experts across five different centers. The method of image capture and collection may differ between these five centers and this factor was not controlled for in the data collection phase. Additionally, the number of images uploaded per human expert per center was not uniform, with some centers and experts contributing more images than others. This process may impact the overall reference standard that ROP.AI was evaluated against, given that human experts were considered the gold standard. The reference standard for this external validation test set was also only graded by a single expert (one of five), and this factor may affect the validation outcome given the know interclinician variability reported in the literature.[5] With the number of images we obtained, we found it to be impractical for all five experts to grade more than 8000 images each; however, each expert remains the sole clinician responsible for ROP screening in their given hospital. We acknowledge the lack of multiple graders as a major limitation to the reference standard. Future studies should formulate a panel of experts who agree on image diagnoses collectively to overcome this interclinician variability.

This study has also allowed us to identify some limitations with the original ROP.AI algorithm, as would be expected for a novel algorithm created from a single center with one expert as the reference standard. The results demonstrate a lower area under the receiver operator curve compared with the original output, with an area under the receiver operator curve of 0.75 for the detection of plus disease compared with 0.99 in the original paper.[15] There are several potential contributing factors to this lower performance, including the lack of image augmentation, different retinal field of view inclusion criteria, inter-grader variability compared with the original expert, and completely new geographic cohort of retinal images. These factors, however, represent realistic conditions encountered in the real-world clinical setting and, although they pose challenges for algorithm performance, would create a superior algorithm with real-world applicability if they can be overcome. In reality, experts are able to review multiple RetCam images for the same eye as well as critical clinical information such as birth weight and gestational age, which may contribute to their decision for diagnosis and follow-up. As a standalone algorithm, ROP.AI is unable to assimilate these clinical conditions; however, further algorithm training could use the analysis of images as a whole (per eye) rather than individual images. This strategy may improve the overall performance of the algorithm. Additionally, further training of the ROP.AI algorithm with this external test set should strengthen its diagnostic ability and may improve the overall performance of ROP.AI with an aim to improve generalizability.

This study demonstrates ROP.AI's potential to be a screening tool for the diagnosis of ROP; however, the overall performance remained low when compared with its development area under the receiver operator curve, sensitivity, specificity, and negative predictive value. This result highlights the importance of external validation and outlines the evidence of both geographical variation and interexpert variability between the New Zealand development image set and the Australian external test set. Our study has uniquely accepted multiple variables that were previously excluded from the training test set such as optic disc location, lower image quality, and lack of augmentation. These features are likely to contribute to the lower performance established; however, they remain important to include to assimilate realistic clinical practice. Further training with these externally collected images from multiple centers across Australia into the existing ROP.AI algorithm should strengthen its diagnostic ability and improve generalizability. The successful advancement of such an algorithm may pave the way toward a fully automated diagnostic system that could revolutionize screening for ROP.

Acknowledgments

The authors thank Isha Gupta (Medical student, Bond University).

Supported by the Children's Hospital Foundation Health Services Research Grant 2020 (Grant reference number 50329)

Disclosure: **A. Bai**, None; **S. Dai**, None; **J. Hung**, None; **A. Kirpalani**, None; **H. Russell**, None; **J. Elder**, None; **S. Shah**, None; **C. Carty**, None; **Z. Tan**, None

References

1. Good WV. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. *Trans Am Ophthalmol Soc*. 2004;102:233–248; discussion 248-50.
2. Chiang MF, Quinn GE, Fielder AR, et al. International Classification of Retinopathy of Prematurity, Third Edition. *Ophthalmology*. 2021;128(10):e51–e68.
3. Tasman W. Multicenter trial of cryotherapy for retinopathy of prematurity. *Arch Ophthalmol*. 1988;106(4):463–464.
4. Mintz-Hittner HA, Kennedy KA, Chuang AZ. Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity. *N Engl J Med*. 2011;364(7):603–615.
5. Gschließer A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *Am J Ophthalmol*. 2015;160(3):553–560.e3.
6. Fleck BW, Williams C, Juszczak E, et al. An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials. *Eye (Lond)*. 2018;32(1):74–80.
7. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. *Early Hum Dev*. 2008;84(2):77–82.
8. Chow SSW, Creighton P, Chambers GM, et al. Report of the Australian and New Zealand Neonatal Network 2017. Sydney: ANZNN; 2019.
9. Yu TY, Donovan T, Armfield N, et al. Retinopathy of prematurity: the high cost of screening regional and remote infants. *Clin Exp Ophthalmol*. 2018;46(6):645–651.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
11. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136(7):803–810.
12. Redd TK, Campbell JP, Brown JM, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. *Br J Ophthalmol*. 2019;103:580–584.
13. Huang YP, Basanta H, Kang EY, et al. Automated detection of early-stage ROP using a deep convolutional neural network. *Br J Ophthalmol*. 2020;105(8):1099–1103.
14. Tong Y, Lu W, Deng QQ, et al. Automated identification of retinopathy of prematurity by image-based deep learning. *Eye Vis (Lond)*. 2020;7:40.
15. Tan Z, Simkin S, Lai C, et al. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. *Transl Vis Sci Technol*. 2019;8(6):23.
16. Bai A, Carty C, Dai S. Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: a systematic review. *Saudi J Ophthalmol*. 2022;36(3):296–307.
17. Chen J, Campbell JP, Ostmo S, Chiang MF. Automated assessment of stage in retinopathy of prematurity using deep learning. *Invest Ophthalmol Vis Sci*. 2020;61(7).
18. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136(7):803–810.
19. Wang J, Ju R, Chen Y, et al. Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine*. 2018;35:361–368.
20. Yildiz VM, Tian P, Yildiz I, et al. Plus disease in retinopathy of prematurity: convolutional neural network performance using a combined neural network and feature extraction approach. *Transl Vis Sci Technol*. 2020;9(2):10.

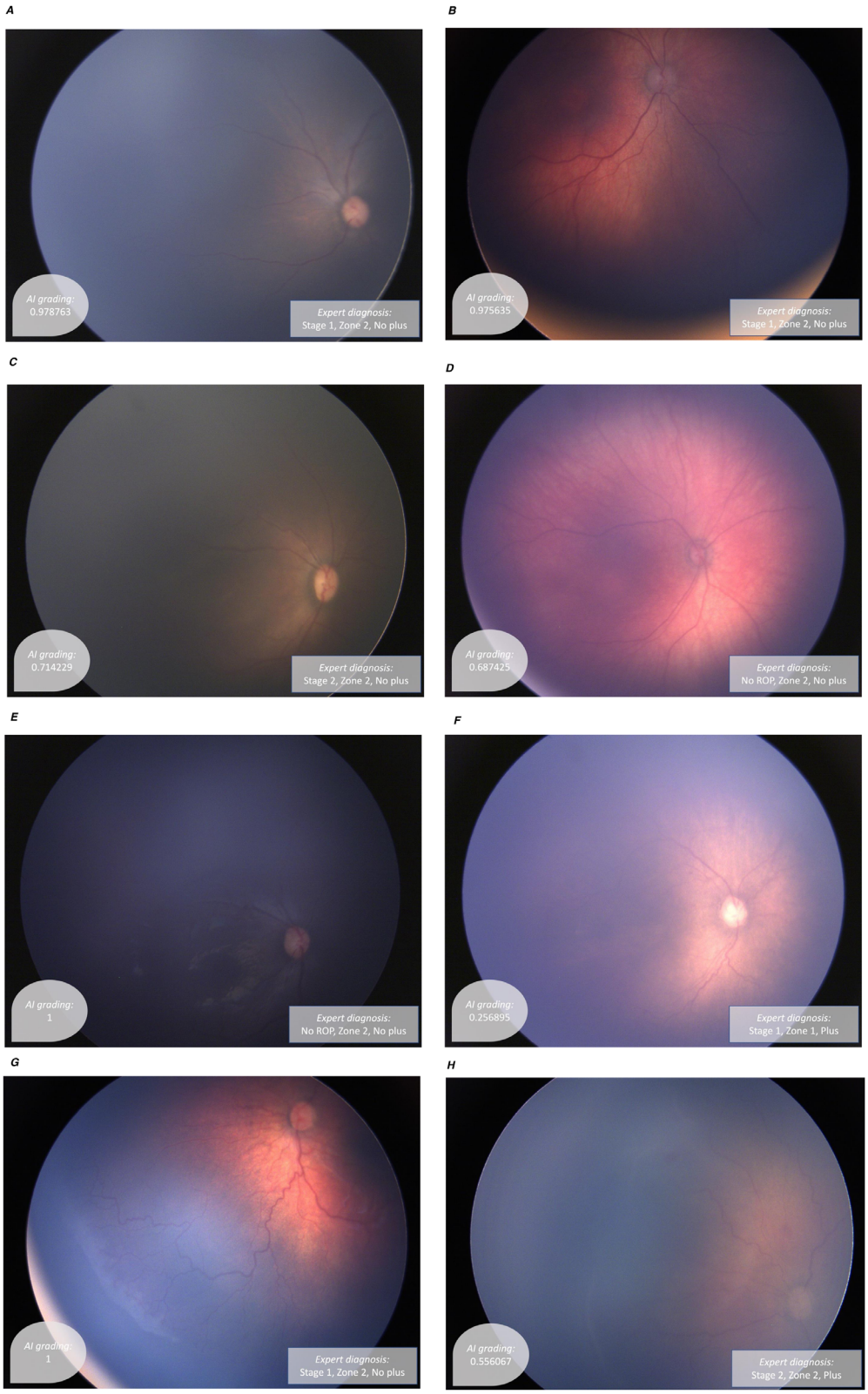
Appendix 1

Screening guidelines for involved hospitals found below. Infants born at less than 31 weeks gestation age or a birth weight of less than 1250 g underwent ROP screening at all four centers.

https://www.health.qld.gov.au/__data/assets/pdf_file/0019/1023553/o-rop.pdf.

Appendix 2

A random selection of 8 Retcam images from the 90 images reviewed for misclassification. Discrepancies seem to occur in images that are slight blurred, of dark fundi, or with the optic nerve near the periphery of the image.


A
AI grading:
0.978763
Expert diagnosis:
Stage 1, Zone 2, No plus
B
AI grading:
0.975635
Expert diagnosis:
Stage 1, Zone 2, No plus
C
AI grading:
0.714229
Expert diagnosis:
Stage 2, Zone 2, No plus
D
AI grading:
0.687425
Expert diagnosis:
No ROP, Zone 2, No plus
E
AI grading:
1
Expert diagnosis:
No ROP, Zone 2, No plus
F
AI grading:
0.256895
Expert diagnosis:
Stage 1, Zone 1, Plus
G
AI grading:
1
Expert diagnosis:
Stage 1, Zone 2, No plus
H
AI grading:
0.556067
Expert diagnosis:
Stage 2, Zone 2, Plus