A New Era in the Screening and Diagnosis of Retinopathy of Prematurity: the Application of Artificial Intelligence

Author

Bai, Amelia

Published

2022-12-15

Thesis Type

Thesis (Masters)

School

School of Pharmacy & Med Sci

DOI

10.25904/1912/4725

A NEW ERA IN THE SCREENING AND DIAGNOSIS OF RETINOPATHY OF PREMATURITY: THE APPLICATION OF ARTIFICIAL INTELLIGENCE

Dr. Amelia Yun Bai
MD, BSc

School of Medical Science
Griffith University

Submitted in fulfilment of the requirements of the degree
Master of Medical Research

July 2022

Synopsis

Retinopathy of prematurity (ROP) is a sight threatening proliferative retinal vascular disease affecting premature infants. Vision loss in ROP is preventable through the early identification and treatment of severe disease. Timely screening and accurate diagnosis is therefore crucial for the diagnosis of ROP, however, multiple challenges exist in current screening processes including limited access to expert ophthalmologists required for ROP screening, subjectivity of diagnosis and cost and time burdens involved in transporting infants to tertiary hospitals. Artificial intelligence (AI) has the potential to overcome current challenges in ROP diagnosis and may transform the way ROP is screened for and managed. Through innovative deep learning technology, a well-designed, well-validated detection algorithm may provide accessible, objective analysis of retinal images to assist expert ophthalmologists in detecting referrable ROP. This thesis will introduce readers to the pathophysiology and grading of ROP, evidence for current treatment guidelines and AI applications in ophthalmology. The systematic review will provide the background evidence into requirements for an accurate, reliable AI algorithm in ROP diagnosis and the validation of our AI algorithm, ROP.AI, will provide insight into the revolutionary diagnostic potential of this deep learning program. Finally, we will discuss future plans for ROP.AI including a methodology proposal to implement the algorithm into a prospective clinical trial for the diagnosis of ROP.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself. There may be some content repetition throughout the thesis due to unmodified articles prepared for publication.

Amelia Bai

TABLE OF CONTENTS

FIGURES

TABLES

Included in this thesis are two papers in Chapter 2 and Chapter 4 which are co-authored with other researchers. My contribution to these papers are outlined at the front of the relevant chapter.

CHAPTER 1: BACKGROUND

RETINOPATHY OF PREMATURITY

Introduction

Retinopathy of prematurity (ROP) is a vasoproliferative retinal disease affecting premature infants. It is characterized by abnormal fibrovascular proliferation in the immature retina and is sight-threatening, affecting 50,000 children worldwide. [1] Those born less than 30 weeks gestation or less than 1250g birth weight are most at risk as preterm birth removes the visual system from the nurturing intrauterine environment. [2] Accurate and early diagnosis is integral to the successful treatment of the preventable blindness caused by this disease. Despite improvements in ophthalmic care, ROP continues to remain one of the leading causes of childhood blindness worldwide. [1] [3] Multiple challenges exist which contribute to this, including the subjective nature of diagnosis, limited access to ROP screening, as well as the increased need for highly specialised paediatric ophthalmologists for diagnosis and treatment.

Pathophysiology

The normal retinal vascular development occurs in two main phases; vasculogenesis and angiogenesis. [4] The first phase, vasculogenesis, occurs where mesenchymal spindle cells emerge from the optic nerve head around 14-15weeks of gestation and grow into the superficial layer of the retina. [5] These cells migrate along the distribution of the future retinal blood vessels out temporally from the optic disc. They do not migrate to the retinal periphery and are not seen after 21 weeks gestation. The second phase of normal retinal vascular development is mediated by angiogenesis, where new blood vessels develop from existing blood vessels. From 17-18 weeks of gestation, networks of capillaries form by angiogenesis and spread towards the leading peripheral edge of the retinal vasculature. From 25 weeks gestation the leading edge of retinal capillaries has multiple arcades of arterial and venous vessels. Angiogenesis appears to be driven by tissue hypoxia, stimulating hypoxia inducible factor -1 (HIF-1) which acts as a transcription factor for Vascular Endothelial Growth Factor (VEGF). [4] The relative hypoxic condition in utero allows HIF-1 degradation to be prolonged. This then intern stimulates VEGF to promote endothelial cell angiogenesis, allowing normal retinal vascular development.

The pathology which occurs in retinopathy of prematurity also occurs in two phases, as it comprises of both reduced retinal vasculature formation and the abnormal proliferation of retinal vasculature. The first phase begins when retinal vascular growth ceases after birth. HIF-1 is rapidly degraded in extra-utero tissue conditions, which leads to reduced angiogenesis of retinal vasculature. [6] Additionally, hyperoxia produced by oxygen therapy exacerbates this effect. The second phase of ROP occurs as the retinal neural cells continue to develop, through proliferation and differentiation causing increased metabolic activity despite the vascular-deplete environment. Eventually, the immature retina becomes critically hypoxic. It is here that pathological VEGF overproduction occurs, causing abnormal angiogenesis and subsequent neovascularisation. [6] Insulin-like growth factor (IGF-1) has also been linked to ROP development, as IGF-1 is required for normal VEGF signalling. [7] Without the normal IGF-1 provided by the placenta and amniotic fluid in-utero, premature infants have lower levels of serum IGF-1 which may contribute further to ROP development. [8] The abnormal growth of blood vessels as a result may lead to extraretinal extension causing traction on the vitreous and possible retinal detachment and blindness.

Classification

The International Classification of Retinopathy of Prematurity [9] is the globally accepted guideline in classifying disease location and severity of ROP. Originally published in 1984, expanded in 1987, revised in 2005, and most recently updated in 2021, this classification system has been the mainstay in further understanding the progression of the disease. The classification system is made up of nine components including zone, extent, and stage of acute disease, as well as the presence of plus or preplus disease, aggressive ROP, regression, reactivation, and long-term sequelae.

1. ***Zone of Vascularization:*** Defines three concentric zones of the retina centred around the optic disc (**Error! Reference source not found.**). The location of the most posterior retinal vascularization indicates the zone for that eye. The radius of the circle represented by Zone I extends from the centre of the optic disc to twice the distance from the centre of the optic disc to the centre of the macula. Zone II consists of a ring shaped region extending nasally from the outer limit of Zone I to the nasal ora serrata and with a similar distance temporally, superiorly, and inferiorly. A new zone termed *posterior zone II* is new to the third edition classification and defines a region of two

disc diameters peripheral to the zone 1 border. Posterior zone II indicates potentially worse disease than ROP in peripheral zone II. Zone III is the residual crescent of the temporal retina anterior to zone II. The term *notch* has additionally been introduced to describe an intrusion of ROP lesion of 1 to 2 clock hours along the horizontal meridian into a more posterior zone than the rest of the disease. If this is present, the disease is recorded as the most posterior zone, followed by "secondary to notch".

Image Removed

FIGURE 1: Right Eye (Re) And Left Eye (Le) Showing Zone Borders And Clock Hours For The Location And Extent Of Retinopathy Of Prematurity. The Dotted Circle Represents Borders Of Posterior Zone Ii, And An Example Of Stage I Disease In Zone Ii Is Depicted In The Le. (Adapted From The International Classification Of Retinopathy Of Prematurity, 3rd Edition 2021[9])

2. ***Extent of Disease***: This defines the circumferential location of disease based on clock hours or 30° sectors.
3. ***Staging of Acute Disease***: Describes the abnormal vascular response at the junction of the vascularized and avascular retina. As a whole, the staging for an eye is determined by the most severe presentation.
    a. ***Stage 1*** – Demarcation Line: Thin line separating the avascular retina anteriorly from the vascular retina posteriorly
    b. ***Stage 2*** – Ridge: A ridge in the region of the demarcation line with both height and width extending above the plane of the retina. Small isolated tufts of neovascular tissue may be found on the surface of the retina, known as 'popcorn'
    c. ***Stage 3*** – Extraretinal Neovascular Proliferation: Neovascularisation which extends from the ridge into the vitreous
4. Retinal Detachment (Stages 4 & 5):
    a. ***Stage 4*** – *Partial Retinal Detachment*: Begins at the point of fibrovascular attachment to the vascularized retina and are generally concave and circumferentially oriented. The extent of the retinal detachment depends on the number of clock hours and degree of fibrovascular traction. Stage 4 is divided into:
        i. ***Stage 4A***: Partial retinal detachment spared fovea
        ii. ***Stage 4B***: Partial retinal detachment involved fovea
    b. ***Stage 5*** – *Total Retinal Detachment*: Usually funnel shaped and caused by traction but can occasionally be due to exudate. Subcategorized into:

i. ***Stage 5A:*** Optic disc is visible by ophthalmoscopy

ii. ***Stage 5B:*** Optic disc is not visible secondary to retrolental fibrovascular tissue or closed-funnel detachment

iii. ***Stage 5C:*** 5B plus anterior segment abnormalities

5. ***Plus and Preplus Disease:*** The terms *preplus* and *plus disease* represent a continuous spectrum of abnormal retinal vascular changes. Plus disease is defined by the appearance of dilation and tortuosity of retinal vessels, as assessed by vessels within zone I.
6. *Figure* 2 is used in the International classification as standard examples of the features of plus disease. *Preplus disease* defines abnormal vascular dilation and tortuosity of the posterior pole vessels that is not severe enough to be deemed plus disease (
7. Figure 3). As the amount of ROP activity is on a continuous spectrum of severity, the vessel abnormalities may progress to plus disease over time.

Image Removed

FIGURE 2: Wide-Angle Fundus Photographs Demonstrating Examples Of Plus Disease. **Above**: Plus Disease With Notable Venous Dilation And Arterial Tortuosity. Note That Plus Disease Is Out Of Proportion To Visible Peripheral Findings, Suggestive Of Flat Neovascularization (Stage 3; White Arrows). **Below**: Severe Plus Disease, With Dilation And Tortuosity Of Both Arteries And Veins. (Adapted From The International Classification Of Retinopathy Of Prematurity 2021[9])

Image Removed

FIGURE 3: Preplus Disease, With Notable Arterial Tortuosity But Minimal Venous Dilation (Adapted From The International Classification Of Retinopathy Of Prematurity 2021[9])

8. ***Aggressive ROP (A-ROP)***: This refers to an uncommon, rapidly progressive, severe form of ROP with characteristic features of prominent plus disease, circumferential extension and an ill-defined nature of retinopathy (
9. Figure 4). Early A-ROP may exhibit capillary abnormalities posterior to the original border of the vascularized retina such as arteriovenous shunting, resembling dilated vascular loops surrounding areas of vascular injury (
10. Figure 5). A-ROP does not progress through the classic stages 1 to 3 and left untreated usually progresses to stage 5 disease. This version of ROP can be diagnosed on a single visit as very early in the development the retinal vessels show increased dilatation and tortuosity in all 4 quadrants out of proportion to the peripheral retinopathy.

Image Removed

FIGURE 4: Fundus Photograph Showing A-ROP With Border Between Vascular And Avascular Retinal In Zone I. White Arrow Shows Dilated Vascular Loops. Note Also Diffuse Flat Extraretinal Neovascularization And Severe Plus Disease (Adapted From The International Classification Of Retinopathy Of Prematurity 2021 [9])

Image Removed

FIGURE 5: A-ROP With Severe Vasoconstriction, Capillary Nonperfusion, Nonphysiologic Dilated Vascular Loops, And Arteriovenous Shunts And Plus Disease In Zone I (Adapted From The International Classification Of Retinopathy Of Prematurity 2021 [9])

11. ***Regression:*** Defines disease involution and resolution, characterised by thinning and whitening of neovascular tissue. Typically starts with vascular changes such as decreased plus disease and vascularization into peripheral avascular retina.
12. ***Reactivation:*** Occurs when there is recurrence of acute phase features and can occur after incomplete or complete regression. It is seen more frequently following anti-VEGF treatment.
13. ***Long-Term Sequelae***: Represents the spectrum of ocular abnormalities that may lead to permanent sequelae including retinal detachments, retinoschisis, persistent avascular retina, glaucoma, and macular anomalies such as blunting or absence of the foveal depression.

Treatment

Strict oxygen weaning and monitoring can prevent the development of ROP due to the role that hyperoxia plays in the pathogenesis of the disease [10]. The incidence of severe ROP requiring treatment has been shown to be lower in babies with a lower oxygen saturation target [11, 12] but is associated with increased infant mortality. [12] Following the publication of The BOOST II, SUPPORT, and COT trials, [12-14] higher (91-95%) saturation targets for extremely preterm neonates have been adopted by all neonatal intensive care units across Australia and New Zealand. [15] This approach has likely further increased the incidence of ROP, resulting in more premature infants requiring ongoing screening for ROP. While most ROP regresses, a small proportion can develop severe visual impairment if left untreated. Randomised control trials have clearly shown that early detection and treatment of severe ROP disease significantly decreases visual impairment and adverse outcomes. [16, 17] Early screening of at risk premature infants is therefore crucial to ROP management and will be discussed in further detail in Chapter 2.

Treatment modalities for ROP include cryotherapy, indirect laser photocoagulation, and more recently injection of intravitreal anti-VEGF. Cryotherapy involves the treatment of avascular retina with a cryoprobe subsequently reducing unfavourable outcomes such as retinal folds or retinal detachments [18]. Indirect laser photocoagulation uses an infra-red diode laser to ablate the areas of avascular retina, reducing the stimulus for new vessel formation resulting in reduced disease progression. The procedure is usually performed under general anaesthetic and can often be labour intensive. The use of intravitreal anti-VEGF agents, specifically Bevacizumab, have been increasingly used off label as an alternative therapy [19, 20]. The prospective randomised control trial Bevacizumab Eliminates the Angiogenic Threat for Retinopathy of Prematurity (BEAT-ROP) [20] compared intravitreal Bevacizumab to conventional laser therapy. They were able to show superior efficacy of intravitreal Bevacizumab for stage 3+ ROP in zone I compared to laser photocoagulation. Additionally, intravitreal Bevacizumab showed significantly lower rates of recurrence of zone I and posterior zone II ROP with continued vessel growth into the peripheral retina, unlike laser photocoagulation [20].

Although the International Classification of Retinopathy of Prematurity [9] provides a detailed classification system for ROP, it does not recommend treatment thresholds for ROP. The cryotherapy of retinopathy of prematurity (CRYO-ROP) study [17] was a multicentre randomised control trial which defined treatment 'threshold disease' as stage 3 ROP in zone I or II in five contiguous or eight non-contiguous clock hours, with plus disease. CRYO-ROP showed that ablation of the avascular retina with transscleral cryotherapy was able to reduce the unfavourable structural and visual outcomes of those infants with threshold disease. This threshold was previously used as the 'cut off' for treatment, however cryotherapy often produces significant periocular inflammation, apnoea, bradycardia, and requires infants to undergo general anaesthesia for the procedure [21]. Since the development of indirect laser photocoagulation, studies have shown that laser coagulation is as effective and may be associated with reduced risk of unfavourable outcomes compared to cryotherapy [21, 22]. Laser photocoagulation has since replaced cryotherapy as the preferred treatment for ROP.

The early treatment for retinopathy of prematurity (ETROP) study subsequently redefined these guidelines, showing efficacy for earlier treatment with laser photocoagulation [16]. The study defined treatment requiring ROP as 'type 1' and observational as 'type 2' ROP. Type 1 ROP represents those in a high-risk, pre-threshold disease group, and should be treated within

48-72hours of identification [16]. Type 2 ROP, a mild pre-threshold disease group, requires close observation and follow-up for monitoring of progression. Type 1 and Type 2 ROP are defined as:

- *Type 1 ROP:*
    a. Any stage of ROP in zone I with plus disease
    b. Stage 3 in zone I without plus disease
    c. Stages 2 or 3 in zone II with plus disease
- *Type 2 ROP:*
    a. Stages 1 or 2 in zone I without plus disease
    b. Stage 3 in zone II without plus disease

ETROP also recommended two weekly follow-up for cases with stage 1 or 2 in zone II, and three weekly follow-up for stages 1 or 2 in zone III [16]. It is evident from the ETROP classification that the presence of plus disease uniquely distinguishes the treatment group from the observation group. The importance of this must be reiterated, as severe ROP resulting in treatment rarely occurs in isolation of the vascular abnormalities seen in pre-plus and plus disease [16], making detection of these disease features vital for the treatment of ROP.

DIAGNOSIS AND SCREENING

Screening Guidelines

Screening is a cornerstone of ROP management. ROP presents in an easily identifiable at-risk population, has a distinct and detectable early disease stage, known natural history, effective treatment if implemented early, and severe visual consequences if diagnosis is missed. There are two main purposes for ROP screening, to identify at risk infants so they may be closely monitored, and to detect infants with severe disease requiring treatment. ROP screening guidelines have high variability across international regions due to the differences in patient demographics and neonatal care standards. [18] The joint statement from the American Academy of Ophthalmology (AAO) and American Academy of Pediatrics (AAP) recommends screening for all infants <1500 grams (g) or birth weight <30 weeks gestational age (GA). [23]

New Zealand have implemented national guidelines where infants born <1250g or <31 weeks GA undergo ROP screening.[2] This is more restrictive compared to the United States and represents one of the least conservative guidelines used worldwide. Despite this, a retrospective review of ROP screening in the country over 22 years revealed sufficient efficacy to capture all at-risk infants of developing sight-threatening ROP with this criteria. [24] In contrast, larger and more mature newborns are developing ROP in middle and low income countries. [25] The mean birth weight of premature infants across eight middle-income countries was >1000g compared to a mean of <800g across the UK, US, and Canada. There was a similar difference in GA, with a range from >26-35 weeks in middle-income countries compared to <26 weeks in first world countries. [26] This is likely due to the large variations in technology and capacity of health resources, resulting in limited care of premature newborns as the survival of smaller and less mature infants increase. [26] These regional variations in neonatal care need to be considered when assigning a screening criterion appropriate to the population. This can be seen in the national guidelines for India, where infants born <34 weeks GA and <1750g birth weight are recommended for screening. [27]

No national guideline currently exists for Australia. Original guidelines proposed by the National Health and Medical Research Council in 1996 recommended screening at <32 weeks or <1500g, which has since been rescinded. [28] Subsequently, the screening criteria varies across both states and individual neonatal units. Queensland has recently released a consensus statement changing their practice from screening <32weeks or <1500g to <31 weeks or <1250g. [29] This criteria is also used by the department of health in Western Australia. [30] New South Wales have adopted a criteria of <30 weeks GA or <1250g, [31] whereas there is no consensus guideline for Victoria. While there are emerging regional guidelines in Australia, a national consensus remains lacking and may be required for streamlined early diagnosis and treatment.

Challenges in diagnosis

Subjectivity of Diagnosis

The ICROP guidelines [9] outlined in chapter 1 provides an international standard for classifying ROP severity. These guidelines require that at least three components are described: the extent of retinal vascularization, extent of abnormality at junction between vascularized and avascular retina, and presence of pre-plus or plus disease.[9] Despite this classification system there is still significant inter-clinician variability in ROP classification.[32] ROP diagnosis is subjective and qualitative based on retinal vessel appearance.[33] Gschließer et al.[32] produced variable levels of both inter- and intra-expert agreement on plus disease diagnosis. Inter-expert agreement, between different ROP experts, was demonstrated as fair (kappa 0.21 – 0.40) for the diagnostic staging of ROP as well as treatment necessity. Average intra-expert agreement between the same experts on two different days was moderate (kappa 0.41-0.6) for the diagnosis of ROP stage and treatment necessity. These results, however, demonstrate that fully reproducible results cannot be produced even from the same examiner, highlighting the subjectivity in interpreting morphological features rather than quantifiable measures. Campbell et al.[33] suggested that diagnostic variability may be due to individual ophthalmologists evaluating particular retinal features differently. In addition, ophthalmologists are classifying pre-plus and plus disease at earlier stages of severity than nine years prior.[34] Moreover, statistically significant geographic variation in plus disease diagnosis has also been reported,[35] further emphasizing the subjectivity of ROP grading.[36]

Time Intensive

Extensive time and training is required for ophthalmologists to perform ROP screening. The skill of binocular indirect ophthalmoscopy is labour intensive.[37] Even with the adoption of wide field digital retinal imaging (WFDRI), ophthalmologists are still required to accurately and expertly evaluate the images captured, in some cases via telemedicine. The results of these screening examinations are critical for determining further management of patients. ROP treatment is time dependent, with a higher risk of blindness if the potential treatment window is missed. [37] Additionally, neonates require repeated exams to avoid missing treatable disease, all of which take up extensive time both for the screening ophthalmologist and the neonatal team.

Demand for Ophthalmologists

There is an increasing demand for ROP screening as the rates of premature births and survival rates of extremely premature and low birth weight infants rise. Most recent data from 2017 demonstrates 8.7% of all newborns were born premature compared to 8.4% in 2010, accounting to a rise of approximately 920 extra premature babies born per year. Approximately 1% of all newborns were born less than 1500g, accounting to about 2996 infants in 2017.[38] There remains limited access to skilled specialists capable of ROP screening and interpretation to keep up with this demand.[39] The most recent Australian government ophthalmology workforce report [40] demonstrates a total of 896 Ophthalmologists working throughout Australia. 15% of those ophthalmologists work in the public sector equating to only 134 ophthalmologists available publicly. The number of paediatric ophthalmologists in Australia is less than 60 in total and those engaged in ROP screening is well below 20.[40] Additionally, regional and rural shortages have been reported to be a problem worldwide.[37] Within the ophthalmology workforce in Australia, only 16% of the 896 ophthalmologists work outside of a metropolitan city, highlighting a further need for access to those in rural and remote regions.[41]

Economic Cost

Yu et al. [42] evaluated the cost of ROP screening for infants born at rural and regional hospitals in Queensland. They were able to determine that the average cost per infant transported to the Royal Brisbane and Women's Hospital was $5110, with transport costs making up a total of 71.6%. [42] The 131 infants studied cost the Queensland health care system a total of $669,413 over a two-year period. [42] This substantial cost is expected to rise with increased numbers of premature infants being discharged to, and cared at, neonatal intensive care units closer to home where ROP screening services are not readily available. Similar cost burdens have been reported elsewhere around the world with estimates of $5078

- $5430 per infant per exam in Ontario, Canada. [43] Additionally, cost-benefit analyses conducted in Mexico and the United States have calculated cost savings of $5556 and $3628 US dollars per child respectively if an ideal ROP screening program is implemented. [44] Further to the potential for significant cost savings, there remains an ongoing need to search for alternative models of ROP screening in rural and regional areas.

Medico-legal Risk

Medico-legal risks of ROP screening present a real problem as a child losing their site due to missed or late screening could lead to the ophthalmologist suffering legal consequences.[45] The American Academy of Ophthalmology's 2006 survey show a decrease in the number of ophthalmologists willing to manage ROP due to concerns regarding legal liabilities associated with missed examinations or misdiagnoses.[46]

Benefits of timely screening

ROP is more prevalent in infants born at extremely low birth weights or low gestational age.[2] More specifically, infants born <30weeks gestational age or with a birth weight <1250g has been adopted as the criteria for ROP screening in Queensland.[2] Repeated screening is often required to track disease progression as blindness may be prevented if features of treatment requiring ROP (Type 1 ROP) are detected early on. For this reason, early screening and detection of ROP is crucial to improve visual outcomes of neonates by allowing timely treatment. Several trials have identified the presence of plus disease to be the most crucial factor in diagnosing severe, treatment requiring disease. [16, 17] As a result, the presence of ROP plus disease, defined by retinal arterial tortuosity and venous dilatation in the posterior retina, remains a reliable indicator for ROP severity.

Present screening tools

Screening for ROP has traditionally been performed by specialist ROP ophthalmologists with binocular indirect ophthalmoscope bedside examination. However, with the introduction of wide-field digital retinal imaging (WFDRI) systems, such as the commercially available RetCam, many centres have moved to image-based ROP screening [2, 37, 47]. With a reported sensitivity of 100% and specificity of 97.9%, [37] digital imaging proves to be a promising technique in facilitating a new ROP screening paradigm. The benefit of RetCam is that non-physician health professionals, who are not specialist ophthalmologists, can be trained to use WFDRI for retinal imaging without the need for an ROP specialist ophthalmologist being present at the bedside. However, it is important to note that a ROP specialist ophthalmologist is still required to view and interpret the images captured by WFDRI for ROP diagnosis. The photographic documentation of ROP using RetCam WFDRI enables a telemedicine approach to ROP screening possible. The use of WFDRI for telemedicine has been reported to be an accurate, reliable, and efficient alternative to indirect ophthalmoscopy in detecting treatment requiring ROP. [37] A review by Fierson and Capone highlighted studies comparing image findings coupled with telemedicine to bedside exams. [48] It was concluded that digital imaging systems and telemedicine are a beneficial adjunct but not an efficient replacement to on-site ophthalmologist examination. [49] A universal consensus for the standard of care for ROP screening is yet to be determined with the availability of the latest technology. There is, however, an increasing body of evidence to support telemedicine WFDRI for ROP screening as an alternate, improved, screening modality. [50, 51]

ARTIFICIAL INTELLIGENCE

Introduction to artificial intelligence

The concept of artificial intelligence dates back to the 1950s, when Alan Turing first discussed how to build and test intelligent machines in the paper "Computing Machinery and Intelligence". [52] However it wasn't until 1956 at the seminal conference *Dartmouth Summer Research Project on Artificial Intelligence* did the host, John McCarthy, officially coin the term 'artificial intelligence'. This conference introduced a computer program designed to mimic the problem solving skills of a human, catalysing the next twenty years of AI research. [53] Since then, computer fundamentals such as storage and processing are no longer limiting the abilities of AI. This is exemplified in its vast application in speech recognition, photo captioning, language translation, robotics, and even self-driving cars. [54-

56] Its advancements have begun to influence the health care industry and is currently at the forefront of research in medical diagnostics. Most prominently, AI has made an impact in radiology where some of the largest medical AI datasets exist. [57, 58] Additionally, AI models such as ResNet have been found to produce a sensitivity and specificity of 98.1% and 72.9% in detecting wrist fractures from X-rays. [59] It is evident that AI is easily applied to visually orientated specialties such as radiology, and ophthalmology is no exception.

In order to understand the practical capabilities of AI in the medical world, it is important to first understand the basics of this transformative technology. AI in itself encompasses an enormous field and is the general term for any form of technology which has problem-solving or decision-making capabilities. One of the many subtypes of AI is machine learning, where a computer algorithm is developed without the explicit encoding of decision-making rules so that learning occurs automatically. A subset of machine learning is deep learning, which can be further divided into supervised learning, where an algorithm is provided with annotated data (termed 'ground truth') to learn from; or unsupervised learning, where the system must classify unlabelled data on its own.

Deep learning is an advanced form of AI which self-learns from large training sets to program itself to perform certain tasks. It relies on multiple processing layers termed 'neural networks' to apply numerous levels of abstraction to input data. [60] A 2D image, for example, comes in the form of an array of pixels where the first layer of representation learns to recognise the presence or absence of edges at certain orientations or locations on the image. The second layer typically detects particular arrangements of edges. The third layer may then assemble these edge patterns into larger combinations that correspond to parts of familiar objects. The subsequent layers would then go on to detect objects as a combination of all these parts. This essentially forms one 'feature map' corresponding to the output of one particular feature of an image. Convolutional neural networks (CNNs) represent the ability to process multiple feature maps, forming an architecture of outputs for each learned feature. Together, these CNNs are able to analyse multiple features of a single image and formulate a particular pattern of recognition.

In practice, a pre-labelled image such as a *normal* retinal image is fed into a deep learning algorithm. The algorithm then uses CNNs to process the image to find features from that image representing a clinical diagnosis. As more normal labelled images are processed by the algorithm, it is able to measure the error between the output score (the algorithm diagnosis) and the desired pattern of scores (the image diagnosis). The machine then modifies its internal adjustable parameters to reduce this error. [60] A typical deep learning system will have hundreds of millions of adjustable parameters. Therefore, the more images you have to train the machine, the smaller the error of its diagnostic output. After training, the performance of the system should be measured on a different set of examples called a test set. This test set measures the program's ability to generalize its performance on new inputs it has never seen before during training. The scalability and ability of CNNs to extract its own relevant features from data is what makes deep learning so powerful, hence allowing the potential to classify and identify disease.

Artificial intelligence in ophthalmology

Deep learning technology has been used with success in other fields of ophthalmology, particularly in diabetic retinopathy. Ramachandran et al. [61] showed that a deep learning algorithm trained against 1685 retinal photos was able to detect referable diabetic retinopathy with sensitivities and specificities close to 80%. These results are comparable to many other groups who have used artificial intelligence to detect referrable diabetic retinopathy including Gulshan et al.,[62] whose deep learning algorithm produced 87-90% sensitivity and 98% specificity. Ting et al [63] achieved clinically acceptable diabetic retinopathy diagnostic ability in an AI system which was externally tested in 11 multi-ethnic cohorts. Training from a dataset of 494,661 retinal images, Ting et al. were also able to produce high accuracy in diagnosing age-related macular degeneration and glaucoma. [63] Further, Zhang et al. [64] conducted a prospective multicentre screening for diabetic retinopathy using an AI algorithm. The study involved 155 centres across China and enrolled 47,269 patients for the algorithm validation, which revealed a sensitivity of 83.3% and specificity of 92.5% in detecting referable diabetic retinopathy.

Following the great promise AI showed in classifying two-dimensional photographs, De Fauw et al. [65] applied deep learning technology to optical coherence tomography (OCT). OCT is a three-dimensional volumetric imaging technique used commonly in ophthalmology assessments. De Fauw et al. were able to develop an AI model that automatically segmented tissue layers, producing referral recommendations that reached or exceeded human experts upon grading sight-threatening retinal diseases. [65] The performance of these deep learning algorithms are extremely promising, however 'real-world' validation is essential to test the performance of these algorithms in clinical scenarios. Again, screening for diabetic retinopathy remains the forefront of AI in ophthalmology with multiple studies having completed prospective evaluations. One such study by He et al. [66] applied a deep learning algorithm to 889 patients screened in a Chinese community hospital clinic. The detection rates of diabetic retinopathy were similar between those of the ophthalmologists (16.1%) and the AI algorithm (16.3%), with 11.4% of patients diagnosed with referrable diabetic retinopathy by ophthalmologists and 11.6% by AI. They were able to produce a sensitivity of 91.18%, specificity of 98.79%, and area under the curve of 0.95 for detecting referrable diabetic retinopathy. Similar results were also published by Ming et al., [67] who produced a sensitivity of 84.6%, specificity of 98%, and area under the curve of 0.913 for referrable diabetic retinopathy in a separate centre in China. These promising findings from the prospective validation of such deep learning algorithms paves the way for AI to be implemented into clinical practice.

Similar can be said for ROP, whose well established screening guidelines makes it perfect for the incorporation of AI algorithms for detection and escalation to treatment. Recently, Brown et al. [68] trained a deep learning algorithm to detect plus disease using a dataset of 5511 retinal images from America. They were able to produce a sensitivity of 93% and specificity of 94% in detecting severe ROP against an independent test of 100 retinal images. This was based on a single reference standard diagnosis based off 3 independent graders, two of which were ophthalmologists. Although the results from this study are promising, the automated nature of a deep learning algorithm means that results can only be as good as its training data. In a disease whose diagnosis is as subjective as ROP, a single reference standard based off only two ophthalmologists does not prove to negate the subjective nature of determining plus disease. Wang et al. [69] developed an automated deep learning system from over 20,795 images collected from a single centre in Chengdu, China. They were able to develop two deep learning systems from this data set, the first being an algorithm for ROP identification (normal or ROP), the second a deep learning algorithm for ROP grading (minor or severe). ROP identification achieved a 96.62% sensitivity and 99.32% specificity, and ROP grading achieved high sensitivities and specificities retrospectively (88.46% and 92.31% respectively), however this was not correlated prospectively with sensitivities of 84.91% and specificities of 96.9%. This result may have occurred due to the potential overlapping features between their classifications of “minor ROP” and “severe ROP”, resulting in an algorithm with poorly defined cut-off points for either diagnosis. Other limitations from this study include the lack of defined inclusion or exclusion criteria for the images used to develop their algorithms, as-well as the absence of any clear screening criteria for newborns undergoing retinal imaging. In addition, the ability to solely identify the presence of ROP is unlikely to be clinically applicable as this does not guide users to a treatment threshold. Given the critical importance of early diagnosis and treatment to prevent blindness in severe ROP, [16] a clear decision threshold such as the presence of plus disease provides a vital cut-off point to determine when a referral for specialist treatment is required. Lastly, despite having a large number of retinal images for algorithm training, Wang et al. have only used images collected from a single centre. Given the known geographic variability in this disease, [18] the likelihood of Wang et al.'s algorithm to be applicable outside of this dataset would be unknown. These two studies mentioned above represent only a subset of many groups who have begun investigating the potential diagnostic power of AI in ROP. Despite the promising results, further evaluation and consideration is required before implementation into routine clinical practise can occur. For this reason, a systematic review analysing AI algorithms in ROP detection was completed.

THESIS OUTLINE AND AIMS

Statement of the Problem

ROP remains the leading cause of preventable childhood blindness worldwide, with an estimated 50,000 children who are blind from ROP. [1, 3] The devastating consequences of untreated ROP has been proven preventable through early disease detection and treatment. [16] Multiple control trials have improved the standard of care of ROP diagnosis and treatment and established guidelines exist to guide expert ophthalmologists in the management of ROP. [9, 16, 20-22, 70] The problem, however, arises from challenges in ROP screening and diagnosis itself as there is a need for highly subspecialized expert ophthalmologists to undertake screening, diagnosis, and treatment. This proves to be extremely time consuming, with repeated exams required for each premature infant. Limited access to these expert ophthalmologists exacerbates the issue of ROP screening provision, especially in rural and regional areas. [24, 42] Additionally, ROP diagnosis remains highly subjective as grading is based on the appearance of certain retinal features. Despite an international standard for classifying ROP severity, [9] significant inter- and intra-clinician variability remains. [32] As the rates of preterm birth and preterm birth survival increases, the burden of ROP screening and diagnosis is subsequently expected to rise. Current screening processes may be inadequate in meeting increasing demands and this calls for a novel strategy to overcome these challenges.

Significance of the Problem

The rising disease burden of ROP may be prevented with a paradigm shift in ROP screening and diagnosis. Lacking in current screening processes is an accessible, objective, accurate tool to assist clinicians with ROP diagnosis. The development of a fully automated screening system to diagnose ROP disease from captured retinal images could revolutionize ROP screening and may provide a solution to present and future challenges. An automated tool of this magnitude, however, must be well designed with robust results and requires vigorous real-world clinical testing before it can be implemented as a screening tool. To the best of our knowledge, a tool like this has not yet been implemented into clinical practice for the diagnosis of ROP. This thesis will provide the foundation for determining the requirements in the design, development, validation, and implementation of an AI deep learning tool, ROP.AI. The incorporation of ROP.AI into telemedicine structures may provide point-of-care testing at the bedside and could be implemented as an adjunct to current screening regimes, potentially improving the visual outcomes of premature infants undergoing ROP screening.

Thesis Aims

(i) To systematically review the evidence for the performance of AI algorithms in the diagnosis and grading of ROP compared to human expert ophthalmologists. To determine, from the systematic review, the important components of an AI screening program including the methodology of developing an algorithm, algorithm training requirements, performance grading of the AI algorithm and methods for validation. (*Chapter 2*)

(ii) To retrospectively collect a database of RetCam (Natus Medical Incorporated) retinal images captured during routine ROP screening from five expert ophthalmologists across five different centres in Australia. In doing so, to develop an efficient, user friendly system for clinicians to upload retinal images into a secure, private database for the collection of the corresponding clinical diagnoses. (*Chapter 3*)

(iii) To validate the deep learning algorithm ROP.AI, trained previously on images collected from a single clinician in New Zealand, on new retinal images collected from across five centres in Australia. To determine the level of generalizability in the existing ROP.AI algorithm through statistical methods and comparing the level of agreement between ROP.AI diagnosis and that of expert ophthalmologists. (*Chapter 4)*

(iv) To discuss prospective research goals for further training and refinement of the existing ROP.AI algorithm, and to explore future directions for AI in ROP screening and diagnosis. (*Chapter 5*)

CHAPTER 2: SYSTEMATIC REVIEW

Performance of deep learning artificial intelligence algorithms in detecting retinopathy of prematurity: a systematic review

Authors: Amelia Bai[1,2,3], Christopher Carty[4,5], Shuan Dai[1,3,6]

Author affiliations:

1. Department of Ophthalmology, Queensland Children's Hospital, Brisbane, Australia
2. Centre for Children's Health Research, Brisbane, Australia
3. School of Medical Science, Griffith University, Gold Coast, Australia
4. Griffith Centre of Biomedical and Rehabilitation Engineering (GCORE), Menzies Health Institute Queensland, Griffith University Gold Coast, Australia
5. Department of Orthopaedics, Children's Health Queensland Hospital and Health Service, Queensland Children's Hospital, Brisbane, Australia
6. University of Queensland, Australia

Citation: 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: p. 296-307.** [DOI: 10.4103/sjopt.sjopt_219_21]

Statement of contribution to co-authored paper

For this co-authored manuscript, the candidate conceived and designed the systematic review, conducted all steps in the review process, analysed all results and prepared the manuscript.

Amelia Bai | 30/7/22 | *Candidate*

A/Prof Shuan Dai | 31/7/22 | *External supervisor and corresponding author*

Abstract

Background: Artificial intelligence (AI) offers considerable promise for retinopathy of prematurity (ROP) screening and diagnosis. The development of deep learning algorithms to detect presence of disease may contribute to sufficient screening, early detection and timely treatment for this preventable blinding disease.

Objective: This review aimed to systematically examine the literature in AI algorithms in detecting ROP. Specifically, we focused on the performance of deep learning algorithms through sensitivity, specificity, and area under the receiver operating curve (AUROC) for both the detection and grade of ROP.

Methods: We searched Medline OVID, PubMed, Web of Science, and Embase for studies published from Jan 1, 2012, to Sep 20, 2021. Studies evaluating the diagnostic performance of deep learning models based on retinal fundus images with expert ophthalmologists' judgement as reference standard were included. Studies which did not investigate the presence or absence of disease were excluded. Risk of bias was assessed using the QUADAS-2 tool.

Results: Twelve studies out of the 175 studies identified were included. Five studies measured the performance of detecting the presence of ROP, and seven studies determined the presence of plus disease. The average AUROC out of 11 studies was 0.98. The average sensitivity and specificity for detecting ROP was 95.72% and 98.15% respectively, and for detecting plus disease was 91.13% and 95.92% respectively.

Conclusions: The diagnostic performance of deep learning algorithms in published studies were high. Few studies presented externally validated results or compared performance to expert human graders. Large scale prospective validation alongside robust study design could improve future studies.

Keywords:

Artificial Intelligence, Deep Learning, Screening, Diagnosis, Retinopathy of Prematurity

Introduction

The concept of artificial intelligence dates back to the 1950s, when Alan Turing first discussed how to build and test intelligent machines in the paper “Computing Machinery and Intelligence”. [52] It wasn’t until 1956, however, at the seminal conference *Dartmouth Summer Research Project on Artificial Intelligence*, did John McCarthy officially coin the term ‘artificial intelligence’ (AI). This conference introduced a computer program designed to mimic the problem solving skills of a human, catalysing the next twenty years of AI research. [53] Today, AI is incorporated into many applications for day to day life, including speech recognition, photo captioning, language translation, robotics, and even self-driving cars. [54-56] These applications are made possible through the use of deep learning, an advanced form of AI which self-learns from large training sets to program itself to perform certain tasks. [60] Application of AI has gained popularity in the medical diagnostic field, and promising outcomes have resulted from deep learning screening algorithms in Ophthalmology.

There has been particular success in AI screening for diabetic retinopathy, with several groups reporting deep learning algorithms detecting diabetic retinopathy at sensitivities and specificities of 83-90% and 92-98% respectively [62] [64]. Moreover, the successful validation of these algorithms has seen progression to 'real-world' implementation of screening programs through prospective evaluation. One such study produced a sensitivity of 83.3% and specificity of 92.5% in detecting referable diabetic retinopathy in a prospective evaluation. [64] Similar promising results are being reported by many other groups utilising deep learning for the diagnosis of other ophthalmic conditions including diabetic macula oedema [65], age-related macular degeneration, [71] glaucoma, [72] and retinopathy of prematurity (ROP). [68, 73]

ROP is a retinal vascular proliferative disease affecting premature infants whose diagnosis is dependent on timely screening. Globally, it is estimated that at least 50,000 children are blind from ROP [1] and it remains the leading cause of preventable childhood blindness. [3] Advances in retinal imaging means disease is now easily identifiable by retinal photographs, making it a perfect candidate for deep learning. As survival rates of premature infants continue to increase with medical advances,[74] the demand for ROP screening is rapidly exceeding the capacity of available specialist ophthalmologists. For this reason, reports of deep learning models matching or exceeding human experts in ROP diagnostic performance has generated considerable interest. It remains fundamental, however, that this enthusiasm does not overrule the need for critical appraisal as a missed diagnosis of ROP can result in significant sequelae such as blindness. Therefore, any deep learning screening algorithm will need to show high diagnostic performance, high sensitivity, be generalisable, and be applicable to the real-world setting. In anticipation of deep learning diagnostic tools becoming implemented into clinical practice, it is judicious to systematically review the body of evidence supporting AI screening for ROP. This systematic review aims to critically appraise the current state of diagnostic performance of deep learning algorithms for ROP screening, with particular consideration for study design, algorithm development, type of validation, performance compared to clinicians, and diagnostic accuracy.

Methods

Search strategy and selection criteria

Studies that developed or validated a deep learning model for the diagnosis of ROP and compared accuracy of algorithm diagnoses to ROP experts were included in this systematic review. We searched MEDLINE-Ovid, Pubmed, Web of Science, and Embase for studies published from Jan 1, 2012, to Sep 20, 2021. The full search strategy for each database is available in *appendix 1*. The cut-off of Jan 1, 2012 was prespecified based on an important breakthrough made with the development of deep learning approaches in the model AlexNet [75]. The search was first performed on July 10, 2020, revised on May 23, 2021 and updated on Sep 20, 2021. Manual searches of bibliographies and citations from included studies were also completed to identify any additional articles potentially missed by searches.

Eligibility assessment was conducted by two reviewers who independently screened titles and abstracts of search results. Only studies aiming to identify through AI algorithms the presence of the disease of interest, ROP, were included. We accepted standard-of-care diagnosis, expert opinion or consensus as adequate reference standards to classify the absence or presence of disease. We excluded studies that did not test for diagnostic performance or investigated accuracy of image segmentation rather than disease classification. Studies which assessed the ability to classify disease severity were accepted if they incorporated primary results of disease detection. Review articles, conference abstracts and studies that presented duplicate data were excluded. We assessed the risk of bias in patient selection, index test, reference standard, and flow and timing of each study using QUADAS-2. [76] Full assessment of bias can be found in *appendix 2*.

This systematic review was completed following the recommendations of the PRISMA [77] statement and the research question was formulated according to the CHARMS [78] checklist for systematic reviews of prediction models. Methods of analysis and inclusion criteria were specified in advance.

Data Analysis

Data was extracted independently by two reviewers (AB and SD) using a predefined data extraction sheet, followed by cross-checking. Any discrepancies were discussed with a third reviewer (CC). Demographics and sample size (gestational age, birth weight, number of participants, number of images), data characteristics (data source, inclusion and exclusion criteria, image augmentation), algorithm development (architecture, transfer learning, number of images for training and tuning), algorithm validation (reference standard, number of experts, same method for assessing reference standard, internal and external validation), and results (sensitivity, specificity, area under the receiver operating characteristic curve for algorithm, human graders, and external validation if applicable) were sought. Two papers produced different algorithms from different data sets or with different identification tasks and were therefore recorded as separate algorithms in the results section. [69, 79] Data from all 12 papers were included and any missing information was recorded. In the case where sensitivity and specificity were not explicitly recorded but could be calculated from a confusion matrix, the calculated results were included.

Results

Our search identified 175 records, of which 99 were screened (*Figure 1*). 30 full-text articles were assessed for eligibility and 12 studies were included in the systematic review. [68, 69, 73, 79-88] 15 studies were excluded due to no test of diagnostic performance, [89-96] no classification task, [49, 97, 98] no internal validation, [80, 99] no AI algorithm, [100] and not based on standard clinical care. [101]

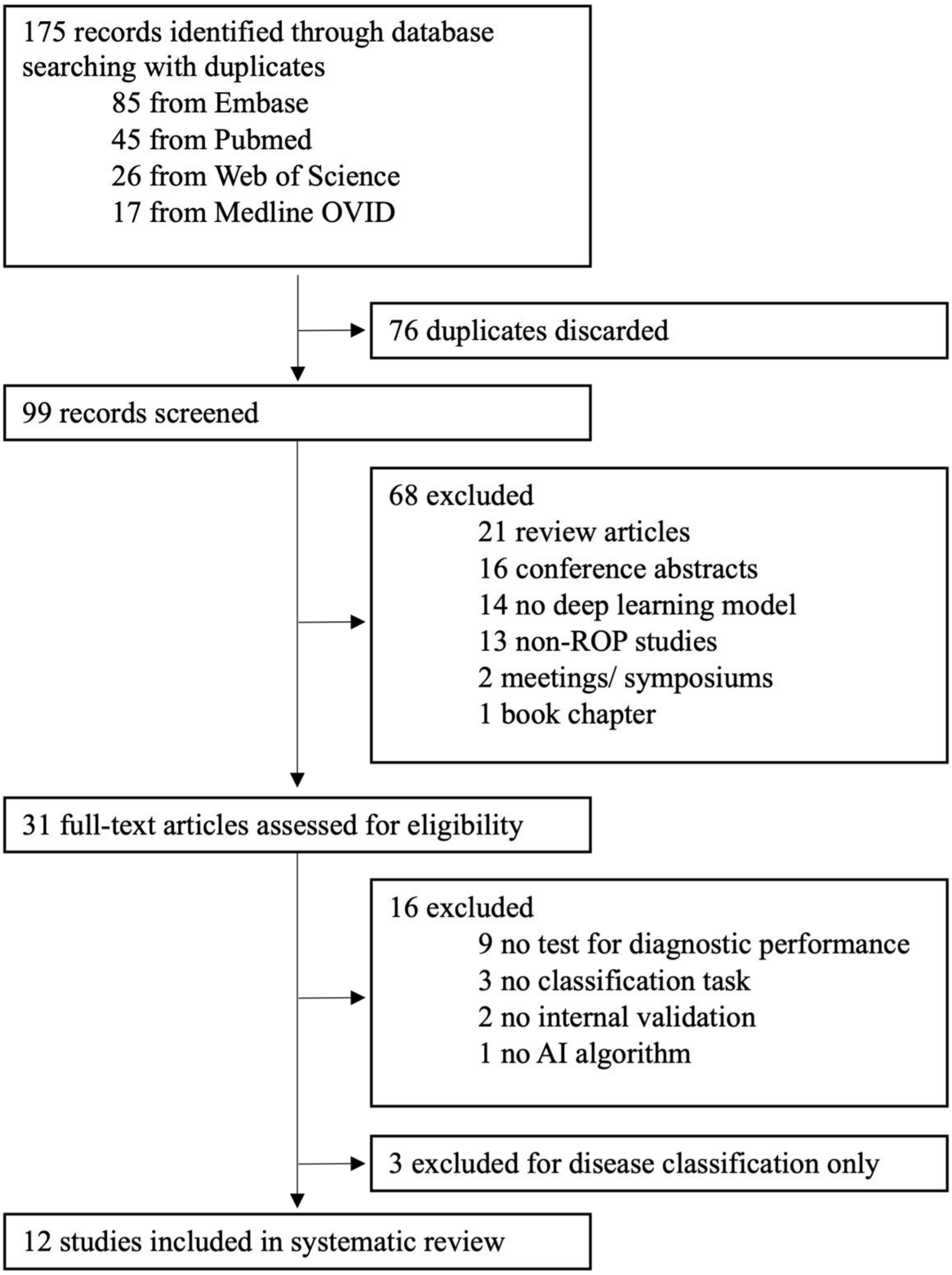


FIGURE 6: Outline Of Study Selection

Data Characteristics and Demographics

All twelve studies obtained retrospective images as part of routine clinical care or from local screening programs. Seven of these studies collected images from China, [69, 81-83, 85, 86, 88] one from India, one from North America, [68] one from American and Mexican sites, [87] one from America and Nepal, [79] and one study included images from New Zealand. [73] Date range for image collection among all studies varied from July 2011 to June 2020. Three studies specified their inclusion criteria [82, 83, 88] and five other studies specified their exclusion criteria. [68, 73, 79, 85, 86] Poor quality images were excluded in five studies [68, 73, 85, 86, 88] and image augmentation occurred in seven studies. [73, 79, 82, 84-87] These characteristics are summarised in Table 1. Seven studies recorded demographic information, [79, 81-84, 86, 88] of which the mean gestational age (GA) was 30.9 weeks and mean birth weight was 1501.25 grams. A total of 178,459 images were used across all 12 studies ranging from 2668 to 52,249 images per study. Five studies formulated an algorithm to detect ROP [69, 79, 81, 82, 88] and seven studies created an algorithm to detect the presence of plus disease out of a total of 5358 plus disease images. [68, 73, 83-87] Full details of demographics and sample size can be found in Table 2.

Data Characteristics

Study	Source of data	Date range	Open-access data	Missing Data	Inclusion Criteria	Exclusion Criteria	Exclusion of poor-quality imaging/ image augmentation
Brown et al 2018	Retrospective cohort, data collected at multiple hospitals across North America	July 2011-December 2016	N	NR	NR	Stage 4-5 ROP	Y / NR
Chen et al 2020							
American Trained Algorithm	Routine ROP screening from 9 North American institutions	NR	NR	NR	NR	Evidence of prior treatment (laser photocoagulation and scars), retinal detachment (stage >4), if artifacts obscures more than 50% of image	N / Y - scipy package in Python, contrast enhancement, green channel extraction, recizing to 224x224 pixels
Nepalese Trained Algorithm	ROP screening program from 4 urban hospitals in Kathmandu, Nepal (Patan Hospital, Kanti Children's Hospital, Paropakar Maternity and Women's Hospital, Tilganga Institute of Ophthalmology)	NR	NR	NR	NR		
Hu et al 2019	Chengdu Women & Children's Central Hospital	2014 - 2017	NR	NR	NR	NR	NR / NR
Huang et al 2020	Neonatal intensive care unit of Chang Gung Memorial Hospital, Linkou, Taiwan	17 Dec 2013 - 24 May 2019	NR	NR	Premature infants with no ROP, stage 1 ROP & stage 2 ROP ROP screening criteria: BW <=1500g, GA <= 32wks, select few infants with BW 1500-2000g or GA >32 weeks with unstable clinical condition	NR	N / Y
Mao et al 2020	Eye Hospital of Wenzhou Medical University	July 2013 - May 2018	NR	NR	Only images of the posterior retina	NR	NR / NR
Ramachandran et al 2021	Karnataka Internet Assisted Diagnosis of Retinopathy of Prematurity (KIDROP) Bangalore, India	NR	NR	NR	NR	NR	N / Y
Tan et al 2019	Auckland Regional Telemedicine ROP (ART-ROP) image library - from four neonatal ICU in Auckland, New Zealand	2006-2015	NR	NR	NR	Poor image quality	Y - not grossly out of focus, not affected by blur / Y
Tong et al 2020	Images collected from Renmin Hopsital of Wuhan University Eye Centre	1 Feb 2012 - 1 Oct 2016	NR	NR	NR	1 poor image quality, 2 imaging artefacts 3 unfocused scans 4 presence of other disease phenotypes (e g retinal haemorrhage)	Y / Y
Wang et al 2021	4 centres in southern China - Joint Shantou International Eye centre of Shantou University & Chinese University of Hong King (JSIEC), Guangdong Women and Children Hospital in Yuexiu branch (Yuexiu) and Panyu branch (Panjyu), and the Sixth Affiliated Hospital of Guangzhou Medical University and Qingyuan People's Hospital	1 Sept 2018 - 24 June 2020	NR	NR	NR	1 nonfundus photos or fundus photos taken by imaging devices other than RetCam 2 infants with other ocular diseases e g congenital cataract, retinoblastoma, or persistent hyperplastic primary vitreous, and 3 any images with disagreeing labels	Y / Y
Wang et al 2018	Images captured during routine clinical ROP Screening from Chengdu Women & Children's Central Hospital	Jan 2018	NR	NR	NR	NR	NR / NR
Yildiz et al 2020	8 study centres: columbia university, university of illinois at chicago, william beaumont Hospital, children's Hospital Los Angeles, Cedars-Sinai Medical centre, university of Miami, Weill Cornell Medical centre and Asociacion para Evitar la Ceguera en Mexico	July 2011 - December 2016	NR	NR	NR	NR	N / Y- resized to 480x640-pixel input images
Zhang et al 2018	From telemedicine ROP trial (Telemed - R), images collected from 30 hospitals in Guangong and Fujian Provinces of China	June 2013	NR	NR	BW <2000g, preterm infants with BW 2000g but severe systemic disorders (as per paediatrician)	NR	Y - highly-blurry images, very dark or bright images, non-fundus photographs / NR

NR = not recorded

TABLE 1: Data Characteristics For The 12 Included Studies

Participants	Sample Size

Study	Mean GA (SD; range), weeks	Birth Weight (SD; range), grams	Mean age at screening (SD), weeks	Number of participants represented by training data	Number of images used in the study	Number of Images with ROP	Number of images with plus disease	Number of images with preplus
Brown et al 2018	NR	NR	NR	898	5511	n/a	172	805
Chen et al 2020								
American Trained Algorithm	26 6 (2 2; NR)	856 2 (293 7; NR)	NR	711	5943	NR	n/a	n/a
Nepalese Trained Algorithm	32 6 (2 8; NR)	1949 6 (495 8; NR)	NR	541	5049	NR	n/a	n/a
Hu et al 2019	32 (NR; NR)	1994 (NR; NR)	NR	720	2668	1184	n/a	n/a
Huang et al 2020	27 3 (1 8; NR)	936 4 (229 8; NR)	NR	NR	10235	1279 557 (stage 1 ROP) 722 (stage 2 ROP)	NR	NR
Mao et al 2020	31 3 (2 1; NR) training set 31 (2; NR) test set	1643 (419 5; NR) training set 1583 3 (401 6; NR) test set	NR	3021	6161	n/a	290	691
Ramachandran et al 2021	32 4 (1 1; NR) no plus 30 9 (1 8; NR) plus	1350 (240; NR) no plus 1280 (226; NR) plus	NR	150	289	n/a	89	n/a
Tan et al 2019	NR	NR	NR	NR	4926 3487 suitable for training	n/a	1638 (post image pre-processing + data augmentation)	0
Tong et al 2020	NR	NR	NR	NR	36231	n/a	3006 2745 (in training dataset) 261 (in test dataset)	NR
Wang et al 2021	32 9 (3 1; NR)	1925 (774; NR)	NR	8652	52249	n/a	NR	NR
Wang et al 2018								
Id-Net	NR*	NR*	NR	1273 total for both Id-Net and Gr-Net 605 for developing data 264 for data for expert comparison 404 data from web	20,795 13,526 (developing) 2361 (expert comparison) 4908 (from web)	6917 5967 (developing) 293 (expert) 657 (data from web)	n/a	n/a
Gr- Net	NR	NR	NR		5089 4139 (developing) 293 (expert comparison), 657 (data from web)	Severe = 2517 [2305 (developing), 120 (expert), 92 (web)] Minor = 2572 [1834 (developing), 173 (expert), 565 (from web)]	n/a	n/a
Yildiz et al 2020	NR	NR	NR	NR	5512	n/a	163	802
Zhang et al 2018	31 9-32 (NR; 24-36 4)	1490-1500 (NR; 630-2000)	NR	NR	17801	8090	n/a	n/a

NR = not recorded; n/a = not applicable; * data represented as bar graph distribution, unable to calculate mean

TABLE 2: Patient Demographics And Sample Size For The 12 Included Studies

Algorithm Development and Validation

Convolutional neural networks formed the basis for algorithms developed in all twelve studies. A variety of algorithms were utilised for transfer learning including ResNet, ImageNet, U-Net, and VGG-16, (*Table 3*) whereas one study did not use a transfer learning approach. [82] The majority of studies used <6,000 images to train their algorithm, however five studies utilised >10,000 images for algorithm development. [69, 82, 85, 86, 88] The reference standard across all twelve studies were based off disease diagnosis by 1-5 expert graders, with an average of 2.6 human graders agreeing upon each image per study. A variety of internal validation methods were recorded, including random split sample validation and cross-validation. (*Table 4*) Five studies [68, 69, 73, 79, 87] obtained external validation of their AI algorithms, of which one study completed a prospective evaluation of algorithm performance. [69]

Algorithm Development

Study	Algorithm name	Algorithm architecture	Transfer learning applied	Number of images for training/tuning
Brown et al 2018	NR	U-Net architecture	Yes	4409 / 1102
Chen et al 2020				
American Trained Algorithm	NR	ImageNet and Pytorch	Yes (from ResNet)	5235 / NR
Nepal Trained Algorithm	NR	ImageNet and Pytorch	Yes (from ResNet)	4802 / NR
Hu et al 2019	NR	ImageNet and TensorFlow (Inception-v2, VGG-16, ResNet-50)	Yes (from ImageNet)	2068 / 300
Huang et al 2020	NR	Tensorflow	No	10235 / 1137
Mao et al 2020	NR	U-Net and DenseNet	Yes (from ImageNet)	5711/ NR
Ramachandran et al 2021	NR	Darknet-53	Yes (from ImageNet)	289/32 (then retrained by 96 images for final model)
Tan et al 2019	ROP AI	TensorFlow's Inception-v3	Yes	80% of 6974/NR
Tong et al 2020	NR	Faster R-CNN + TensorFlow	Yes (from ResNet)	90% of 26459 / 10% of 26459
Wang et al 2021	J-PROP	NR*	Yes (from Res-Unet)	75% (39029) /10% (5140)
Wang et al 2018	DeepROP	Tensorflow - Inception-BN Network	Yes (from ImageNet)	17665 / NR
Yildiz et al 2020	I-ROP ASSIST	NR*	Yes (from U-Net)	5512 / NR
Zhang et al 2018	CAD-R (computer-aided diagnosis system for ROP)	NR*	Yes (from VGG-16)	17801 / NR

NR = not recorded
* specific architecture not recorded, however convolutional neural networks used

TABLE 3: Details Of Algorithm Development For The 12 Included Studies

Algorithm validation

Study	Reference Standard	If compared to experts, how many?	Same method for assessing reference standard across samples	Type of Internal Validation	Number of images for internal validation	External validation	Number of images for external validation
Brown et al 2018	Expert consensus	3	Yes	Random split sample validation	20%	Yes	100
Chen et al 2020							
American Trained Algorithm	Expert consensus	3	Yes	5-fold cross-validation	10% test set	Yes	247 images from Nepal
Nepal Trained Algorithm		1	Yes	5-fold cross-validation	10% test set	Yes	708 images from America
Hu et al 2019	Expert consensus	3	Yes	Random split sample validation	300	No	n/a
Huang et al 2020	Expert consensus	3	Yes	5-fold cross-validation	244	No	n/a
Mao et al 2020	Clinical diagnosis by one ophthalmologist	1	NR	Random split	450	No	n/a
Ramachandran et al 2021	Expert consensus	3	Yes	80:20 split	161 (67 ROP)	No	n/a
Tan et al 2019	Expert ophthalmologist from New Zealand; External images graded by expert ophthalmologist from HongKong	1	No - 2 different experts between internal and external validation	80:20 random split validation	20% of 6974	Yes	90 (33 plus, 57 normal) + additional 26 pre-plus imagesfor assessing pre-plus
Tong et al 2020	Expert grading (11 retinal experts for 1st-round screening, 2 senior experts confirmed or corrected labels)	2	No - 11 different first round graders	10-fold cross-validation	9772	No*	n/a
Wang et al 2021	Expert grading (2 junior ophthalmologists labelled, any disagreement submitted to 1 senior ophthalmologist)	3	No - dependent on agreement	Random split 75:10:15 (training, validation, test) - but based on a patient-based split policy (i.e. all images of a patient were allocated into the same sub-data set)	8080	No	n/a
Wang et al 2018	Expert consensus (images included if 2 out of 3 graders agreed, disagreements sent to fourth ophthalmologist)	3 to 4	Yes	Random split	298 (for Id-Net), 104 (for Gr-Net)	Yes - prospective evaluation	2361 (total, Id and Gr net)
Yildiz et al 2020	Expert Consensus	3	Yes	5-fold cross-validation	5000	Yes	100 (15 plus, 34 preplus)
Zhang et al 2018	Cross validation by one senior ophthalmologist	5 (2 senior experts, 2 attending physicians, 1 resident)	Yes	Random selection	1742 (155 ROP, 1587 without ROP)	No	n/a

* No external validation, however did measure algorithm vs human graders on anotherr 1227 images collected during routine clinical care

TABLE 4: Method Of Algorithm Validation For The 12 Included Studies

Algorithm Performance

The performance of each algorithm is listed in *Table 5*. Five studies recorded the ability of their algorithm to detect the presence of ROP disease, of which the average area under the receiver operating characteristic (ROC) curve (AUROC) was 0.984. [69, 79, 81, 82, 88] Sensitivity and specificity were recorded in four of those studies and averaged 95.72% and 98.15% respectively. [69, 81, 82, 88] One study compared human grader performance to the AI algorithm revealing similar sensitivities (94.1% AI, 93.5% human) and specificities (99.3% AI, 99.5% human) of ROP diagnostic performance. [88] Two of the five studies underwent external validation revealing an average sensitivity and specificity of 60% and 88.3% respectively for detecting presence of disease. [69, 79] The seven other studies determined ability of their algorithm to detect the presence of plus disease. Among these, six studies measured AUROC, with which the average was 0.98. [68, 73, 83, 84, 86, 87] The average sensitivity and specificity for detecting plus disease recorded from six studies was 91.13% and 95.92% respectively. [68, 73, 83-86] External validation occurred in two of these studies and produced an average sensitivity of 93.45% and specificity of 87.35%. [68, 73] Performance of AI algorithm at detecting pre-plus disease was measured in two articles producing an average sensitivity of 96.2% and specificity of 95.7%. [68, 83] This is compared to four studies who measured performance of determining the stage of ROP disease, showing an average sensitivity and specificity of 89.07% and 94.63% respectively. [69, 82, 85, 86]

		Algorithm Performance						Human Performance					External Validation					
	Study	Sens %	Spec %	Area under the ROC curve (AUROC)				Sens %	Spec %	Sens grading %	Spec grading %	AUROC	Sens %	Spec %	AUROC			
Detecting Disease	Hu et al 2019	96	98	0 9922				n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a			
	Zhang et al 2018	94 1	99 3	0 998				93 5	99 5	n/a	n/a	n/a	n/a	n/a	n/a			
	Chen et al 2020																	
	American Trained Algorithm	NR	NR	0 99				n/a	n/a	n/a	n/a	n/a	52	99	0 96			
	Nepal Trained Algorithm	NR	NR	0 96				n/a	n/a	n/a	n/a	n/a	44	69	0 62			
	Combined (American & Nepal) Trained Algorithm	NR	NR	0 99				n/a	n/a	n/a	n/a	n/a	98 / 82 (against American / Nepal set)	96 / 99 (against American / Nepal set)	0 99 / 0 98 (against American / Nepal set)			
					Sens grading %	Spec grading %	AUROC grading									Sens grading %	Spec grading %	AUROC grading
Detecting Disease & Stage	Huang et al 2020	96 14 +/- 0 87	95 95 +/- 0 48	0 96	91 82 +/- 2 03 (stage 1)	94 5 +/- 0 71 (stage 1)	0 93 (stage 1)	NR	NR	NR	NR	NR	n/a	n/a	n/a	n/a	n/a	n/a
	Wang et al 2018																	
	Id-Net	96 64	99 33	0 995	n/a	n/a	n/a	NR	NR	NR	NR	NR	84 91	96 9	NR	n/a	n/a	n/a
	Gr-Net	n/a	n/a	n/a	88 46 (minor vs severe)	92 31 (minor vs severe)	0 951 (minor vs severe)	NR	NR	NR	NR	NR	n/a	n/a	n/a	93 33 (minor vs severe)	73 63 (minor vs severe)	NR
					Sens Pre-Plus %	Spec Pre-plus %	AUROC Pre-plus									Sens Pre-Plus %	Spec Pre-plus %	AUROC Pre-plus
Detecting Plus Disease	Brown et al 2018	93	94	0 98	100	94	NR	n/a	n/a	n/a	n/a	n/a	93	94	NR	100	94	NR
	Mao et al 2020	95 1	97 8	0 99	92 4	97 4	NR	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Ramachandran et al 2021	99	98	0 9947	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	Tan et al 2019	96 6	98	0 993	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	93 9	80 7	NR	81 4	80 7	0 977
	Yildiz et al 2020	NR	NR	0 94	NR	NR	0 88	n/a	n/a	n/a	n/a	n/a	NR	NR	0 99	NR	NR	0 97
Detecting Plus & Severity	Wang et al 2021	91 8	97	0 983	98 2 (stage)	98 5 (stage)	0 998 (stage)	100 (compared to J-PROP on same dataset: 100)	99 8 (compared to J-PROP on same dataset: 98 4)	91 7 (stage) (compared to J-PROP on same dataset: 97 9)	99 1 (stage) (compared to J-PROP on same dataset: 97 4)	NR	n/a	n/a	n/a	n/a	n/a	n/a
	Tong et al 2020	71 3	90 7	NR	77 8 ("normal" "mild" "semi-urgent" "urgent")	93 2 ("normal" "mild" "semi-urgent" "urgent")	NR	74 8 (expert 1), 65 9 (expert 2) (for grading "normal" "mild" "semi-urgent" "urgent")	93 4 (expert 1), 92 3 (expert 2) (for grading "normal" "mild" "semi-urgent" "urgent")	n/a	n/a	NR	n/a	n/a	n/a	n/a	n/a	n/a

TABLE 5: Summary Of Results From The 12 Included Studies

Discussion

We found that deep learning algorithms for ROP screening demonstrated sensitivity and specificity metrics that were comparable to neural network algorithms in diabetic retinopathy. [102] Although this estimate supports application potential for deep learning algorithms to be implemented as a real-world diagnostic tool there are several methodological deficiencies that were common across included studies which need to be considered. These include the quality of reference standard, use of sample size calculations, external validation, definition of presence or absence of disease, and the need for prospective evaluation.

Firstly, we found variability in specific algorithm diagnostic targets with the 12 papers split between diagnosing the presence of ROP as a whole versus the presence of plus disease. It is important to differentiate these diagnostic targets as the clinical implication of the findings will differ. Additionally, most studies utilised a reference standard graded by on average 2-3 experts with only one study producing a reference standard diagnosed by 5 clinicians per image. [88] It is well reported that there is a significant amount of intergrader variability in ROP diagnosis due to its subjective nature, [32, 33] therefore caution needs to be taken in recognising the potential for grader bias in studies utilising only a few expert graders.

Secondly, there was a large variety in the number of images used to train each algorithm, ranging from 289 [84] to 39,029 images. [86] Convolutional neural networks learn by computing the error between the machine's output and the image diagnosis; hence the more images used to train a machine the smaller the error of its diagnostic output. [60] For this reason the studies that had sample sizes in the ten-thousands were likely to have more reliable results than those that were trained off hundreds or thousands of images. Nonetheless, no studies reported formal sample size calculations to ensure sufficient sizing of studies. Despite the challenge of sample size calculations in the context of AI algorithms it remains a principal component of any study design and only one paper reported sample size as a limitation in their study. [82] Future studies should consider formulating sample size calculations in order to justify the number of images required for algorithm design.

Thirdly, exclusion of poor-quality images or image augmentation may impact the performance of these deep learning algorithms in the real-world clinical setting. This is a factor which may limit the diagnostic performance of an algorithm as high quality images correlates to high quality diagnoses and smaller algorithm errors. [60] For this reason it is understandable that most papers will exclude poor quality images, however it is important to keep this within reason. Quality of images used to train an algorithm should correspond to the quality of images taken in the clinical setting so that algorithm performance may equate to its real-life performance. It is also for this reason that external validation of an algorithm using an image set outside of the training image set is crucial to determine the generalisability of a study. Only five of the twelve studies completed external validation of which all but one study, showing equivalent performance, revealed inferior algorithm performance compared to their test set. This finding highlights the need for an out-of-sample external validation in these screening algorithms to better understand how the algorithm will perform in the clinical setting.

Fourth, the ground truth or reference standard labels were mostly derived from data collected for other purposes such as a database of ROP images or retrospective routine clinical care notes. Although there exists an internationally accepted guideline for defining presence and stage of ROP, the International Classification of Retinopathy of Prematurity revisited (ICROP) [70] (more recently updated in a 2021 version [9]), only five studies specifically mentioned the ICROP in their methods for defining the reference standard. As ICROP acts as the universally adopted diagnostic criteria for grading ROP it is safe to assume that the other seven studies also used these guidelines, however the criteria for the presence or absence of disease should always be clearly defined in AI studies.

Finally, only one study completed prospective evaluation of their algorithm, a process that is vital to assess the performance on real-world implications. The majority of studies assessed deep learning diagnostic accuracy in isolation, without external validation as mentioned earlier or comparison to experts. Only three studies provided a comparison of AI performance with human performance, allowing for evaluation of real-world application. Without comparison of AI to human performance, the results from the other seven studies are limited in their ability to be extrapolated into health-care delivery. In order for a deep learning diagnostic tool to be applicable in clinical bedside screening it must perform better or comparable to the gold standard, in this case expert diagnosis. More work is required to validate the performance of AI algorithms in comparison to human graders, ideally using the same external test dataset.

It is clear from this systematic review that there still lacks a well-designed randomised head-to-head comparison of an effectively externally validated AI algorithm to human performance in real-time. A study of this magnitude could reveal the possible clinical implications for implementing an algorithm in the clinical setting. For this reason, prospective evaluations of these deep learning diagnostic tests are crucial in order to unveil the bounding potential of AI in both diagnostic and therapeutic medicine. We recognize that there is a large 'black box' issue in deep learning, where image features learned by an algorithm is unknown to the user. [60] It is for this reason that many clinicians are sceptical to entrust clinical care to AI, especially when the clinical features clinicians are familiar with may not be the same features used by an algorithm. This further emphasises the need for well executed studies that minimise bias and are thoroughly and transparently reported. Most of the concerns we have highlighted in this review are avoidable with robust design and it remains critical that these AI diagnostic tests are evaluated in the context of its intended clinical pathway.

Conclusion

AI has been heralded as a revolutionary technology for many industries, and certainly deep learning algorithms for diagnosis of ROP is no exception. Despite the issues we have highlighted in this systematic review, the performance of the twelve deep learning algorithms evaluated have been extremely high, with all studies delivering a recordable AUROC above or equivalent to 0.94. These results outline the ability for AI algorithms to perform comparable to or exceeding human experts, and provides the groundwork for future large-scale prospective studies. Although there are clear screening and treatment guidelines, ROP disease burden continues to rise as increased survival of preterm infants coincides with advancements in medical care. [3] The inadequate accessibility and number of experienced ophthalmologists continues to limit ROP screening and diagnosis. Consequently, the burden of ROP visual impairment is expected to increase unless a novel strategy such as deep learning diagnostic algorithms becomes available. There is no doubt that the successful application of AI in ROP will revolutionalize disease diagnosis through its high predictive performance and streamlined efficiency. The clinical implications of this successful implementation into real-world clinical practice is immeasurable, with translation into high accessibility, high quality, timely screening and the significant reduction in cost of screening. AI will therefore become ubiquitous and indispensable for ROP screening, and it is important that high quality research continues to aid the translation of this transformative technology in order to reduce the incidence of visual loss and blindness from this preventable disease.

CHAPTER 3: METHODOLOGY

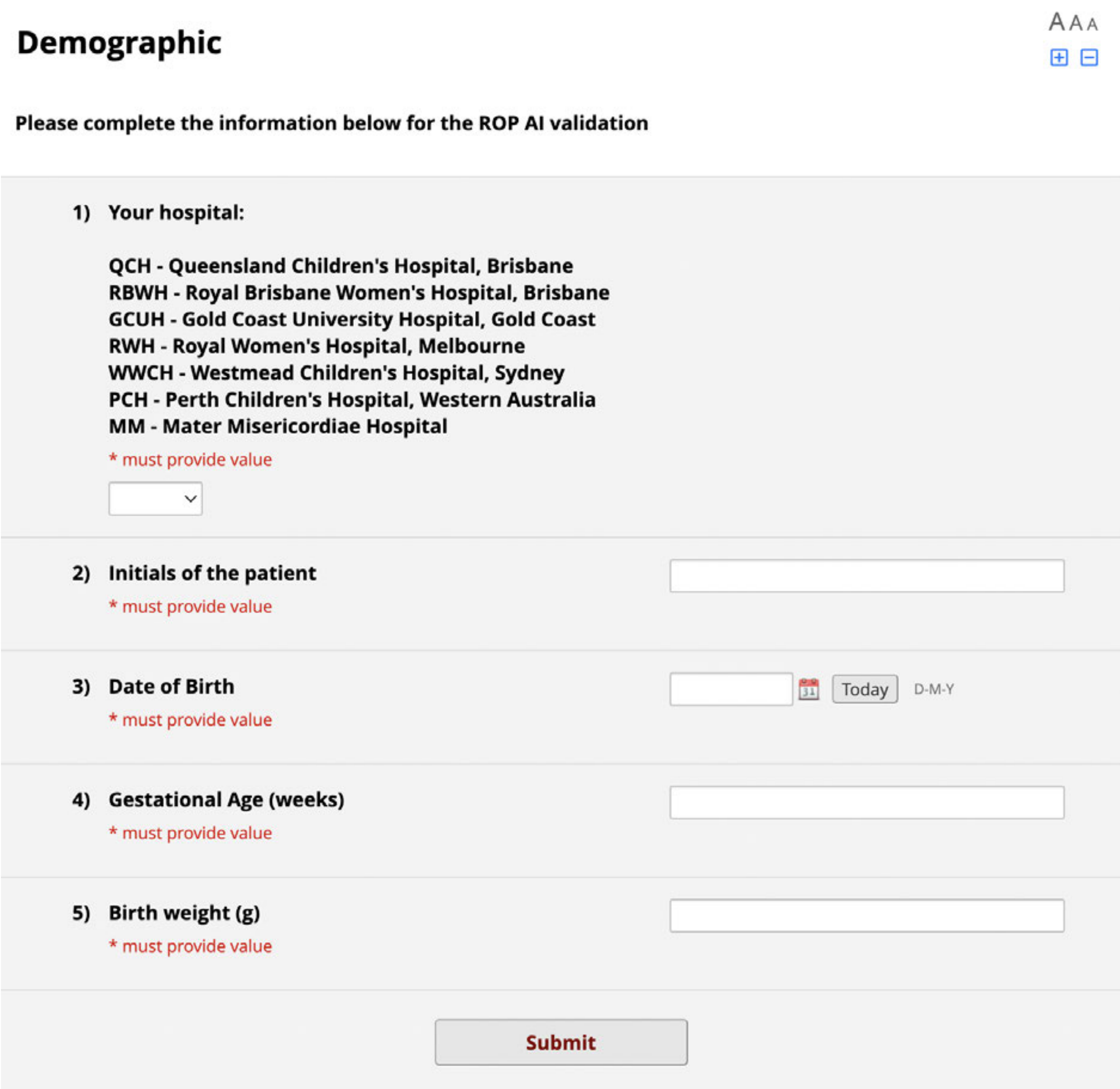
ETHICS AND GOVERNANCE

Ethics approval was sought through the Children's Health Queensland Hospital and Health Service Human Research Ethics committee before the commencement of our research. Approval was granted for the retrospective collection of retinal images from the following sites: Queensland Children's Hospital, Royal Brisbane and Women's Hospital, Gold Coast University Hospital, Royal Children's Hospital Melbourne, Mater Misercordiae, Perth Children's Hospital, and Westmead Children's Hospital. Approval was also granted for the same sites for the prospective validation of ROP.AI through real-time data collection alongside routine clinical screening. HREC Reference Number: **HREC/20/QCHQ/62358**. See ***Appendix 3*** for approval documents.

Site specific agreements were submitted and approved for the Queensland Children's Hospital, Royal Brisbane and Women's Hospital, Gold Coast University Hospital, Mater Misercordiae, and the Royal Children's Hospital Melbourne. The approval letters are present in ***Appendix 4***. A public health act agreement was also sought and approved and all directors of each corresponding department at each of the listed hospitals approved the commencement of our project. Lastly, a research collaboration agreement was formulated between Children's Health Queensland Hospital and Health Service and Griffith University.

DEVELOPMENT OF RETINAL IMAGE DATABASE

A secure software platform was required to upload de-identified RetCam images for the labelling of data. We utilised REDcap (https://www.project-redcap.org/), a private Queensland Health infrastructure capable of providing a storage database of images and data accessible only to those with a login and specific link. A survey was created on REDCap allowing for investigators to input de-identified clinical information to attach to the corresponding retinal images being uploaded. *Figure 7* outlines the first page of the REDCap survey, where hospital site, patient initials, date of birth, gestational age, and birth weight were requested fields. A unique identifier for that patient was automatically created using a combination of their location, initials, and date of birth so that clinical details were untraceable. This assured patient confidentiality and allowed retinal images to be de-identified.



Demographic

Please complete the information below for the ROP AI validation

1) **Your hospital:**

QCH - Queensland Children's Hospital, Brisbane
RBWH - Royal Brisbane Women's Hospital, Brisbane
GCUH - Gold Coast University Hospital, Gold Coast
RWH - Royal Women's Hospital, Melbourne
WWCH - Westmead Children's Hospital, Sydney
PCH - Perth Children's Hospital, Western Australia
MM - Mater Misericordiae Hospital

* must provide value

2) **Initials of the patient**

* must provide value

3) **Date of Birth** Today D-M-Y

* must provide value

4) **Gestational Age (weeks)**

* must provide value

5) **Birth weight (g)**

* must provide value

Submit

FIGURE 7: First page of the REDcap survey requesting for the demographics of the patient whose RETcam images were uploaded.

The second page of the REDcap survey, shown in *Figure 8*, requested collaborators to upload retinal images and their corresponding diagnoses per eye. This was formulated as ROP diagnoses can vary between two eyes of the same patient, and maintaining that differentiation was important. Date of ROP screen or RETCam image capture was also requested so that patients with repeated ROP screening sessions could have multiple entries under the same unique identifier. Stage, zone, and presence of plus or pre-plus disease was also a required field. Collaborators could then choose to upload all images for one eye in a single ZIP file or individually (up to ten) in their original format (jpeg). At the bottom of the survey was a submit button which took the collaborator back to the first page so another patient’s data and images could be uploaded.

The link for the REDCap survey was sent to all collaborators listed in the HREC approval whose hospital had a site specific approval. The author of this thesis uploaded the majority of images for the Queensland Children’s Hospital, Royal Brisbane and Women’s Hospital, and Mater Misericordiae. Only three of the collaborators (Amelia Bai, Shuan Dai, and Jacky Hung) had access to the images and data that was uploaded to REDCap through a secure login.

Date of Screen
* must provide value

Today D-M-Y

Age at Screening (days)

Right Eye

What stage of ROP is this eye?
* must provide value

No ROP 1 2 3 4 5
reset

What is the zone of disease?

1 2 3
reset

Plus disease for the right eye

No plus
Pre-plus disease
Plus disease
reset

Please select how you would like to upload images of the right eye.

Images are in ONE zip file
Images are in Individual files (up to 10 can be uploaded)
reset

Upload a zip file of all the right eye images here

Upload file

Left Eye

What stage of ROP is this eye?
* must provide value

No ROP 1 2 3 4 5
reset

What is the zone of disease?

1 2 3
reset

Plus disease for the left eye

No plus
Pre-plus disease
Plus disease
reset

Please select how you would like to upload images of the left eye.

Images are in ONE zip file
Images are in Individual files (up to 10 can be uploaded)
reset

When you click the Submit button below, you will be taken to a new page to submit the next patient. If you have completed your upload, please close the new window.

Thank you

Submit

FIGURE 8: Second page of the redcap survey outlining the important clinical parameters that were collected as well as the image upload process. Collaborators had a choice of uploading all images for one eye in a zip file or individually.

DATA PROCESSING THROUGH A DEEP LEARNING PLATFORM

Extrapolation of data from REDCap

All Retcam retinal images and their corresponding clinical diagnoses were downloaded from REDCap. Image files were downloaded with the same format of *Record_ID_eye_originalfilename*; where *Record_ID* represented the unique identifier for a specific patient, and *_eye* identified the right or left eye and *_originalfilename* was the original RETCam image file name that was uploaded. This filename convention allowed us to link the record_ID in REDCap to the specific assessment visit for each patient and the specific clinical diagnosis of each eye. Any images that were uploaded in a zip file were first downloaded as a whole, unzipped, and saved in the same format of *Record_ID_eye_originalfilename*. If any records had the same file name, we added a prefix so that every image had its own unique label which corresponded to the recorded diagnosis. File names and their clinical diagnoses (including stage, zone, and plus disease status) were then exported into an excel document.

Image upload into Medicmind

The encrypted AI software platform MedicMind (https://ai.medicmind.tech), which was used to train the previously established deep learning algorithm ROP.AI, was used for the web-based evaluation of ROP.AI performance. MedicMind is a deep learning platform utilising tools from Google's TensorFlow and Inception technology to develop AI algorithms. The ROP.AI algorithm accessible through MedicMind was available to a single collaborator, Jacky Hung. All retinal images collected from REDCap were uploaded in batches into MedicMind for the deep learning algorithm to evaluate. *Figure 9* shows the upload process where images were dragged into the platform and an output between 0 and 1 was produced by the AI. This output represented ROP.AI's diagnosis of the image with 0 representing a normal retina and 1 representing plus disease. Any value in between this continuous scale determined the probability of the image being normal or plus disease. Each image that was uploaded also produced a saliency map (*Figure 10*) outlining which features of the image ROP.AI identified as disease activity. In the example in *Figure 10*, the ROP.AI algorithm produced a probability of 1, indicating a diagnosis of plus disease.

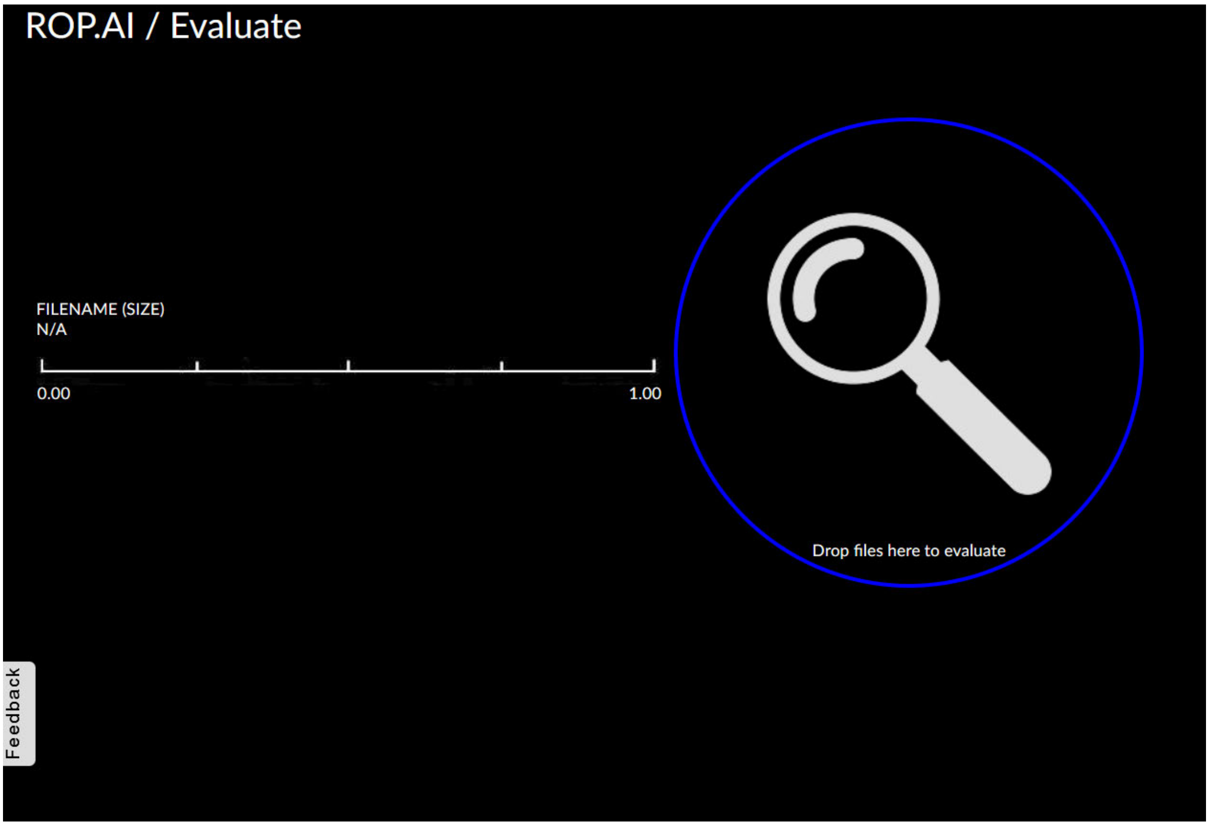


FIGURE 9: Image upload page on the encrypted AI software platform Medicmind. The scale on the left represents the diagnostic output each image produced after being analysed by the ROP.AI algorithm.

FIGURE 10: Evaluation page of ROP.AI, showing an image uploaded on the right with the AI's calculated diagnosis. An output saliency map is also produced on the bottom left.

Data Processing

Results from the ROP.AI evaluation were downloaded from MedicMind into an excel document. Results had the following fields: unique ID from MedicMind, time and date of evaluation, image file name, algorithm name, and value from the evaluation. The image file name from MedicMind remained identical to the image label exported from REDCap so that the corresponding expert ophthalmologist diagnosis could be directly compared. Data from MedicMind was aligned with the data from REDCap to attribute each image with a ROP.AI diagnosis and actual diagnosis. The expert ophthalmologist diagnosis that was uploaded to REDCap was considered as the gold standard.

The collated data (from REDCap and MedicMind) was then simplified into two columns; ROP.AI diagnosis and expert diagnosis. This table was imported into SPSS to perform ROC analysis for the detection of pre-plus and plus disease and corresponding AUROC values were determined. The ROP.AI algorithm produced an automatic cut-off value of 0.5 to distinguish normal from plus-disease images. A confusion matrix was calculated at this cut-off for the detection of plus disease (*Table 6*) and pre-plus and plus combined. (*Table 7*) As the algorithm was initially analysed for the detection of plus disease only, we did not determine its ability to detect pre-plus disease as its own entity. Post-hoc analysis revealed potential ability for ROP.AI to detect pre-plus disease and therefore an analysis for pre-plus and plus combined was created. Further results from the validation of ROP.AI performance is outlined in Chapter 4.

ROP.AI Diagnosis

		Plus Disease	*No Disease*	*Total*
Clinician Diagnosis	*Plus Disease*	326	84	410
	No Disease	2771	3784	6555
	Total	3097	3868	

TABLE 6: Confusion matrix for ROP.AI detection of plus disease

ROP.AI Diagnosis

Clinician Diagnosis	Pre-Plus & Plus	No Disease	Total
Pre-Plus & Plus	901	309	1210
No Disease	2196	3559	5755
Total	3097	3868	

TABLE 7: Confusion matrix for ROP.AI detection of pre-plus and plus disease combined

ANALYSIS AND OPERATING POINT OPTIMISATION

The ROP.AI algorithm had a default operating point threshold of 0.5 to determine the presence of plus disease. This threshold, however, could be tuned to match certain requirements for diagnosis as an algorithm can have multiple operating points. For example, if the threshold value was moved to a higher point at 0.7, any value above this point would indicate plus disease and a higher proportion of images would be considered as normal. In contrast, reducing the operating point would increase the amount of false positives however a higher number of true positives would be captured. The sequelae of missed diagnoses in ROP is vision threatening and therefore a screening algorithm should be developed to have a lower operating point threshold. This would produce higher sensitivities and negative predictive values, with resulting compromise to specificity and positive predictive value. Thus, the ROP.AI algorithm was tested at 0.01 operating point intervals between 0.3 – 0.7 to determine an optimal cut-off point. (*Table 8*) A threshold of 0.34 was chosen to enhance sensitivity and negative predictive value and the results are discussed in Chapter 4.

Operating Point	*0.3*	*0.31*	*0.32*	*0.33*	*0.34*	*0.35*	*0.36*	*0.37*	*0.38*	*0.39*	*0.4*	*0.41*	*0.42*	*0.43*	*0.44*	*0.45*	*0.46*	*0.47*	*0.48*	*0.49*	*0.5*
	Plus + pre-plus vs Normal																				
Accuracy	48%	50%	50%	51%	52%	53%	54%	55%	55%	56%	57%	58%	59%	59%	60%	61%	61%	62%	63%	63%	64%
Sensitivity	**89%**	**88%**	**87%**	**87%**	**86%**	**86%**	**85%**	**84%**	**83%**	**82%**	**81%**	**80%**	**80%**	**79%**	**78%**	**78%**	**77%**	**76%**	**76%**	**75%**	**74%**
Specificity	40%	42%	43%	44%	45%	46%	47%	48%	50%	51%	52%	53%	54%	55%	56%	57%	58%	59%	60%	61%	62%
Negative Predictive Value	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**94%**	**93%**	**93%**	**93%**	**93%**	**93%**	**93%**	**93%**	**93%**	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**
Positive Predictive Value	24%	24%	24%	24%	25%	25%	25%	25%	26%	26%	26%	26%	27%	27%	27%	28%	28%	28%	28%	29%	29%
	Plus Disease vs Normal																				
Accuracy	40%	41%	42%	43%	44%	45%	46%	47%	48%	49%	50%	51%	52%	53%	54%	55%	56%	57%	57%	58%	59%
Sensitivity	**91%**	**91%**	**90%**	**90%**	**90%**	**89%**	**88%**	**87%**	**86%**	**85%**	**84%**	**83%**	**83%**	**83%**	**82%**	**82%**	**81%**	**81%**	**81%**	**80%**	**80%**
Specificity	37%	38%	39%	40%	41%	43%	44%	45%	46%	47%	48%	49%	50%	51%	52%	53%	54%	55%	56%	57%	58%
Negative Predictive Value	**99%**	**99%**	**99%**	**99%**	**99%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**
Positive Predictive Value	8%	8%	8%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	10%	10%	10%	10%	10%	10%	10%	11%

Operating Point	*0.51*	*0.52*	*0.53*	*0.54*	*0.55*	*0.56*	*0.57*	*0.58*	*0.59*	*0.6*	*0.61*	*0.62*	*0.63*	*0.64*	*0.65*	*0.66*	*0.67*	*0.68*	*0.69*	*0.7*
	Plus + pre-plus vs Normal																			
Accuracy	64%	65%	66%	66%	67%	67%	68%	69%	69%	70%	70%	71%	72%	72%	73%	73%	74%	74%	74%	75%
Sensitivity	**74%**	**73%**	**72%**	**71%**	**71%**	**70%**	**69%**	**68%**	**68%**	**67%**	**66%**	**66%**	**65%**	**65%**	**64%**	**64%**	**64%**	**62%**	**62%**	**61%**
Specificity	62%	63%	64%	65%	66%	67%	68%	69%	70%	70%	71%	72%	73%	74%	74%	75%	76%	77%	77%	78%
Negative Predictive Value	**92%**	**92%**	**92%**	**92%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**91%**	**90%**
Positive Predictive Value	29%	30%	30%	30%	30%	31%	31%	32%	32%	32%	33%	33%	34%	34%	34%	35%	36%	36%	36%	37%
	Plus Disease s no plus																			
Accuracy	60%	60%	61%	62%	63%	64%	65%	65%	66%	67%	68%	68%	69%	70%	70%	71%	72%	72%	73%	74%
Sensitivity	**79%**	**79%**	**79%**	**78%**	**78%**	**77%**	**77%**	**76%**	**76%**	**76%**	**75%**	**75%**	**75%**	**75%**	**74%**	**74%**	**74%**	**73%**	**73%**	**72%**
Specificity	58%	59%	60%	61%	62%	63%	64%	65%	65%	66%	67%	68%	69%	70%	70%	71%	72%	72%	73%	74%
Negative Predictive Value	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**	**98%**
Positive Predictive Value	11%	11%	11%	11%	11%	12%	12%	12%	12%	12%	13%	13%	13%	13%	13%	14%	14%	14%	15%	15%

TABLE 8: Accuracy, sensitivity, specificity, negative predictive value and positive predictive value of rop.ai at different operating points from 0.3 to 0

CHAPTER 4: ROP.AI VALIDATION

Multi-centre validation of deep learning algorithm ROP.AI for the automated diagnosis of plus disease in retinopathy of prematurity

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

Author affiliations:

1. Griffith University, Gold Coast, QLD, Australia
2. Centre for Children's Health Research, South Brisbane, QLD, Australia
3. Queensland Children's Hospital, South Brisbane, QLD, Australia
4. University of Queensland, St Lucia, QLD, Australia
5. Gold Coast University Hospital, Southport, QLD, Australia
6. Bond University, Robina, QLD, Australia
7. Royal Children's Hospital, Parkville, VIC, Australia
8. University of Melbourne, Parkville, VIC, Australia
9. Mater Misericordiae, South Brisbane, QLD, Australia
10. Menzies Health Institute Queensland, Griffith University, QLD, Australia
11. Aegis Ventures, New York, United States of America

Citation: Bai, A., Dai, S., Hung, J., Russell, H., Kirpalani, A., Elder, J., Shah, S., Carty, C., Smith, B., Tan, Z. (2022). Multi-centre validation of deep learning algorithm ROP.AI for the automated diagnosis of plus disease in retinopathy of prematurity.
[Manuscript prepared for submission]

Statement of contribution to co-authored paper

For this co-authored manuscript, the candidate contributed most to the conception and design of the study, development of data collection methods, analysis of results and prepared the manuscript in consultation with co-authors.

Amelia Bai | 30/7/22 | *Candidate*

A/Prof Shuan Dai | 31/7/22 | *External supervisor and corresponding author*

Introduction

A crucial feature in recognising treatment-requiring ROP is the presence of plus disease, defined as dilation and tortuosity of retinal vessels in the posterior retina. [9] Landmark studies have established that severe ROP through the early detection of plus disease may be effectively treated with cryotherapy, laser photocoagulation, or intravitreal injections of anti-vascular endothelial growth factor [16, 17, 20]. It is therefore essential that the screening and diagnosis of plus disease is 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, [9] significant intra- and inter-clinician variability remains. [32] Moreover, significant geographic variation in plus disease diagnosis has also been reported emphasizing the inconsistency in ROP diagnosis. [35] Secondly, ROP screening is labour intensive and requires extensive training and experience before an ophthalmologist can become proficient at screening. 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. [3] Despite the implementation of wide-field digital retinal imaging systems, such as the commercially available RetCam, ophthalmologists are still required to evaluate multiple images per eye per patient to determine an accurate diagnosis. As a result, the demand for ROP screening has risen to an extent where access to expert ophthalmologists capable of examination and diagnosis is limited. [39] Finally, the shortage of ophthalmologists in regional and rural centres means higher healthcare expenses to provide screening for infants from these areas. The high cost of transporting preterm infants to tertiary centres adds to the logistical difficulties of screening and contributes to the financial burden on health care centres and affected families. [42]

These challenges have incentivised 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 due to its ability to self-learn from training image sets to perform certain tasks. [60] 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 minimise the error. Thus, the larger the number of images used to train the machine, the smaller the error of its diagnostic output. Since AI is self-taught, there lies a major 'black-box' issue, where image features recognised by an algorithm is unknown to the user. [60] For this reason there is hesitancy among clinicians to entrust screening and diagnosis to an AI algorithm. Therefore, AI studies intended to be utilised 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 precisely validate the performance of an algorithm to determine the generalisability into clinical practice.

Groups in the Unites States [68, 103] and China [82, 85] have developed automated artificial intelligence (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 externally validated against a geographically novel data set to ensure reproducible results in a separate population. The deep learning algorithm, ROP.AI, developed from retinal images collected from New Zealand, has been able to achieve very high sensitivity, specificity and accuracy in detecting plus disease compared to an expert ophthalmologist (96.6%, 98%, and 97.3% respectively). [73] This system, however, has been trained based off one expert ophthalmologist's grading of retinal images and the generalisability of ROP.AI's performance to a geographically new population is unknown as it has been trained on New Zealand infants only. The aim of this study is to investigate the performance of ROP.AI against retinal images collected from 5 different centres across Australia and compare their diagnostic performance in detecting plus disease with 5 expert ophthalmologists as human graders.

Methods

A total of 6945 retinal images taken with RetCam (Natus Medical Incorporated) were retrospectively collected and deidentified across 5 different centres in Australia from May 2021 to February 2022. All images that included the optic disc in its field of view were accepted and only images that were unreadable for clinician diagnoses were excluded. Images were not edited or adjusted and did not undergo augmentation. The corresponding clinician diagnosis for each image was also collected and images were labelled as either normal, pre-plus, or plus disease. Diagnoses were supplied by one of 5 experts corresponding to their real-world clinical diagnosis during routine ROP screening.

The performance of the ROP.AI algorithm, through the cloud-based platform MedicMind (https://ai.medicmind.tech), was evaluated against all 6945 images. The ROP.AI diagnosis for each image 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 (ROC). ROP.AI performance in diagnosing plus disease vs no plus disease, and plus disease with pre-plus disease vs normal was compared to the reference standard diagnosis.

ROP.AI produced a probability value between 0 and 1 following evaluation of retinal images. A default threshold of above 0.5 was initially utilised to determine an image with plus disease. Given the prospect ROP.AI holds to be utilised as a screening algorithm, this operating point of 0.5 was optimised to produce high sensitivity and negative predictive value to reduce the likelihood of missed diagnoses. The algorithm was retested at 0.01 operating point intervals between 0.3 and 0.7 until an optimal point of 0.34 was identified. The sensitivity, specificity, and negative predictive value was re-calculated for this cut-off point.

Results

A total of 6945 images from 885 individual eyes were analysed, of which 5755 (82.87%) images were normal, 800 (11.52%) images had pre-plus disease, and 390 (5.61%) images had plus disease. The New Zealand trained ROP.AI algorithm produced an area under the ROC of 0.79 for detection of plus disease (*Figure 8*) and 0.76 for detection of pre-plus and plus disease combined. (*Figure 9*)

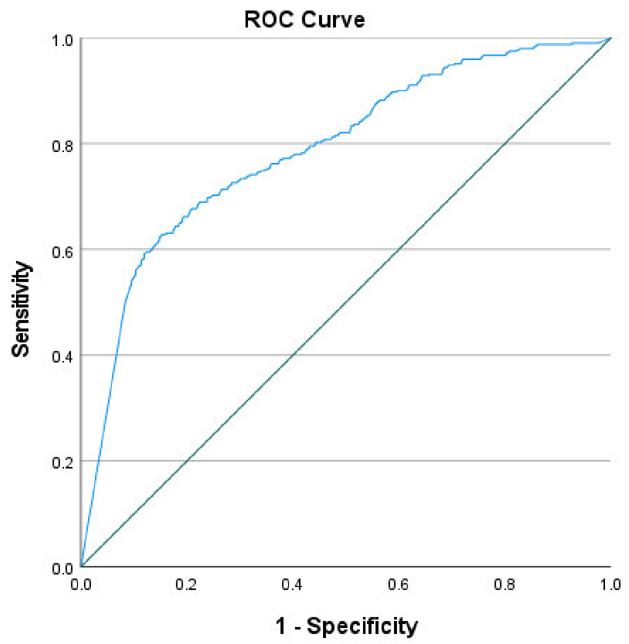


FIGURE 11: ROC Curve For ROP.AI Diagnosis Of Plus Disease

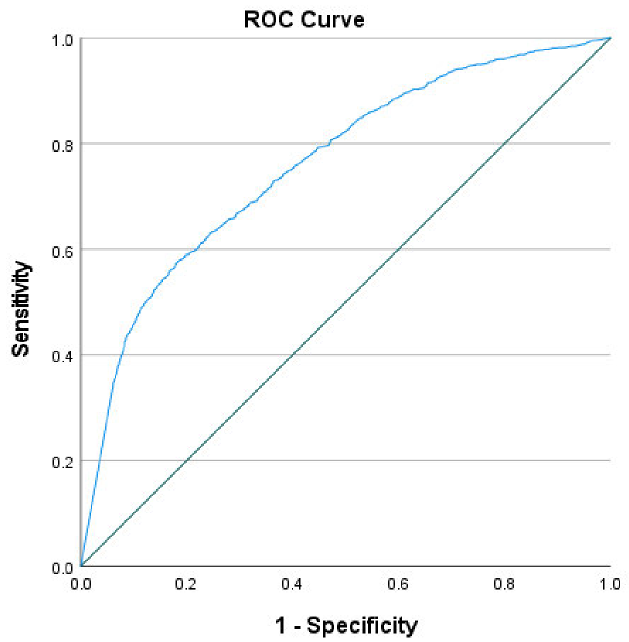


FIGURE 12: ROC curve for ROP.AI diagnosis of pre-plus and plus disease combined

Sensitivity, specificity, and negative predictive value at the algorithm default cut-off point of 0.5 were 80%, 58%, and 98% respectively in detecting plus disease. Sensitivity, specificity, and negative predictive value in detecting pre-plus and plus disease combined were 74%, 62%, and 92% respectively. The average outputs produced by the algorithm for normal, pre-plus, and plus disease images were 0.43 (SD 0.3) , 0.7 (SD 0.3), and 0.8 (SD 0.29) respectively. The distributions of these probability outputs are illustrated in *Figure 10*.

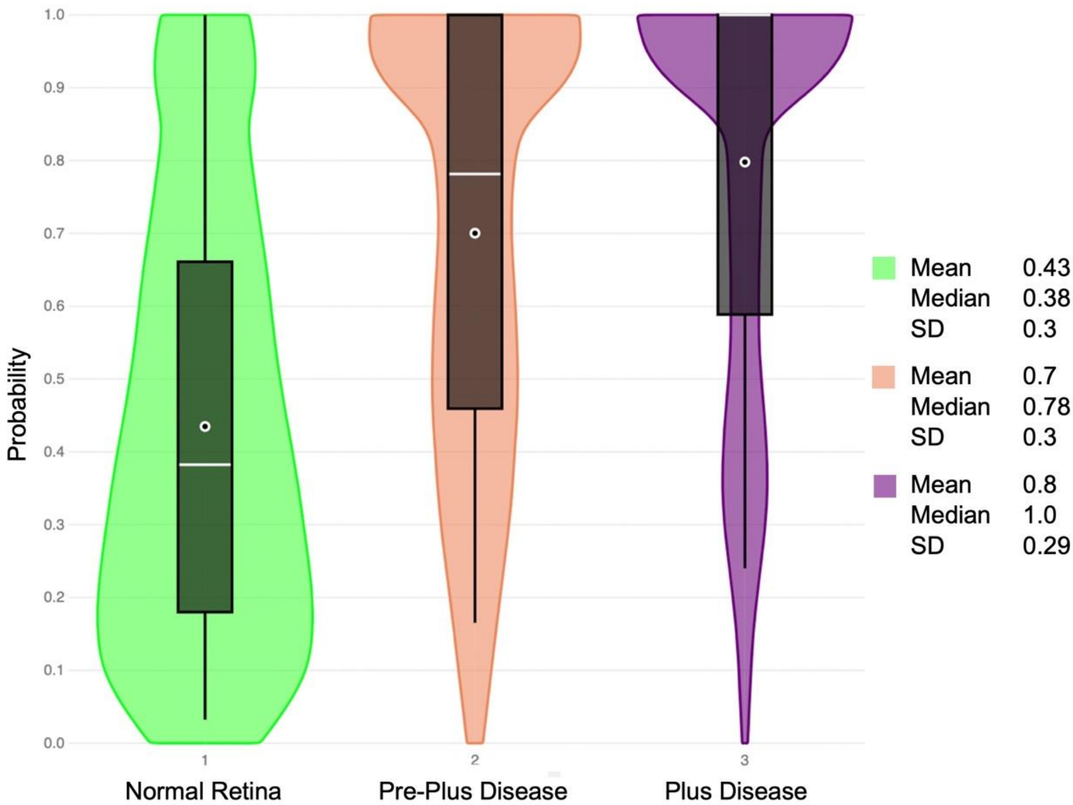


FIGURE 13: Violin Plot For The Distribution Of Algorithm Performance In Normal, Pre-Plus, And Plus Disease Fundal Images.

Following operating point optimisation to 0.34, sensitivity, specificity and negative predictive value improved to 90%, 40%, and 99% for the detection of plus disease and 86%, 45%, and 94% respectively for the detection of pre-plus and plus disease combined.

Discussion

This study evaluated 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 centres across Australia. We found that the original ROP.AI algorithm performs comparatively well against these new images, supporting its potential for application into real-world clinical use. The key findings were (1) ROP.AI demonstrates generalisability with high performance in the diagnosis of 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 or worse and (3) ROP.AI produced high performance in diagnosing images without pre-processing or augmentation, assimilating realistic clinical practise 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. [9] Despite this, inconsistency in ROP and plus disease diagnosis remains prominent, [32] and statistically significant geographic variation in plus disease diagnosis has also been reported. [35] For this reason, an objective screening system such as ROP.AI holds potential to reduce clinically significant management differences and improve outcomes for premature infants. However, it is crucial that the algorithm can both overcome geographic variability and inter-expert variability. Our study is unique in that it provides a large test set of over 6000 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 centres across Australia). To the best of our knowledge, this is the first automated deep learning algorithm to have undergone large-scale external validation in ROP. Our results reveal high ROP.AI performance with a sensitivity and negative predictive value of 90% and 99% in the diagnosis of plus disease at the optimal cut-off point 0.34. Although ROP.AI was only trained to recognise plus disease, the algorithm has been able to detect pre-plus disease or worse with a sensitivity and negative predictive value of 86% and 94%. Despite being trained from a single centre with only one clinician as the reference standard, ROP.AI's performance on this external test set has been comparable to those reported by other groups with a much smaller number of external images. [69, 87, 104, 105] These findings support ROP.AI's potential of reproducibility into the clinical setting with high diagnostic output despite analysing images it has never been exposed to before.

In addition, the images used to validate ROP.AI were not preselected, augmented, or pre-processed 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 as it is well recognised among AI that high quality images correlates to high quality diagnoses and smaller algorithm errors [60]. Meticulous exclusion of poor-quality images and small 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 quality and scope of images corresponding to those taken in the clinical setting so that the validation of our algorithm performance may equate to its real-life performance. Preliminary retrospective review of a small subset of images which had the highest disagreement between experts and ROP.AI grading revealed many were of dark or slightly blurred photos. These factors may have contributed to the low specificities we obtained (40% for the diagnosis of plus disease, 45% for pre-plus or worse). In a disease such as ROP which holds devastating sight-threatening consequences if treatment is delayed, high sensitivity and negative predictive value are most crucial to avoid missed diagnoses. As a result, low specificity may occur as a compromise, as it does here following optimisation 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 reduce health care costs despite low specificities. Further algorithm training, however, may overcome this and should be considered prior to implementation into clinical practice.

This study has several limitations to consider. Firstly, images were collected as part of routine clinical screening by five human experts across five different centres. Method of image capture and collection may differ between these five centres and these were not controlled for in the data collection phase. Additionally, the number of images uploaded per human expert per centre was not uniform with some centres and experts contributing more images than others. This may impact the overall reference standard that ROP.AI was evaluated against given that human experts were considered the gold standard. Future studies should formulate a panel of experts who agree upon image diagnoses collectively to overcome the inter-clinician variability reported in the literature. [32] This is especially important as the calculation of other outcome measures such as determining image 'gradeability' is extremely limited given the subjectivity of ROP diagnoses. Finally, thisstudy has allowed us to identify some limitations with the original ROP.AI algorithm, as would be expected for a novel algorithm created from a single centre with one expert as the reference standard. The results demonstrate a lower AUROC compared to the original output, with an AUROC of 0.79 for the detection of plus disease compared to 0.99 in our original paper. [73] There are several potential contributing factors to this lower performance, including lack of image augmentation, different retinal field of view inclusion criteria, inter-grader variability compared to the original expert, and complete new geographic cohort of retinal images.

These factors, however, represent realistic conditions encountered in the real-world clinical setting and, although pose challenges for algorithm performance, would create a superior algorithm with real-world applicability if they can be overcome. This study has aimed to validate the ROP.AI algorithm against these realistic clinical conditions and has produced promising results. Further training of the ROP.AI algorithm with this novel test set should strengthen its diagnostic ability and generalisability and may pave the way towards a fully automated diagnostic system which could revolutionize ROP screening.

Conclusion

This study demonstrates ROP.AI's potential to be a generalizable screening test for the diagnosis of ROP, with high sensitivities and negative predictive values in the detection of both plus disease and pre-plus disease. It highlights the importance of external validation and outlines the evidence of both geographical variation and inter-expert variability between New-Zealand test sets and Australian test sets. Future training of these images collected from multiple centres across Australia into the existing ROP.AI algorithm should strengthen its diagnostic ability and will be published in future studies.

CHAPTER 5: THE FUTURE OF ROP.AI

Impact of the validated ROP.AI

This study has successfully validated the use of ROP.AI, trained from a single clinician's data set at a single centre in New Zealand, to multiple clinician diagnoses in various locations around Australia. The strong inter- and intra-clinician variability reported in the literature demonstrates a major challenge for ROP diagnosis in both clinical and AI settings. [32, 106, 107] The subjectivity of ROP severity and these diagnostic inconsistencies are therefore expected to impact any deep learning algorithm trained off such variable input data. Additionally, training using a single clinician dataset may limit the performance and generalizability of an algorithm by exposing it to potential biases including both selection and information bias. Moreover, AI generally produces higher quality diagnostic outputs with smaller algorithm errors when only high-quality images are utilized in the test set. [60] Our study, however, has limited the exclusion of images to precisely correlate to the quality of images taken in the clinical setting so that our results reflect the algorithm's real-world performance. For this reason, the validation of the ROP.AI algorithm has produced promising results despite being tested against data from a diverse population with four new expert ophthalmologist graders, with no pre-processing or augmentation of images. The high sensitivity (90% at the optimised cut-off point) and negative predictive value (99% at the optimal cut-off point) achieved in diagnosing plus disease demonstrates ROP.AI as a robust algorithm with high performance despite expected inter-user variability and geographic variation in the new dataset.

Upcoming algorithm development

The performance of ROP.AI has been encouraging despite being trained from a single centre with one clinician as the reference standard diagnosis, however, further research efforts are underway to improve the algorithm to a clinical standard. As outlined in our systematic review, a large scale, prospective evaluation of an effectively externally validated AI algorithm is crucial to investigate the generalizability of an algorithm and its clinical implementations. Our study aims to assess the performance of ROP.AI in real-time, comparing head-to-head against current clinical screening. To accomplish this, our future studies will bring together the five leading expert ophthalmologists from across Australia to create a research consortium that will drive the evolution of ROP.AI. (Figure 6)

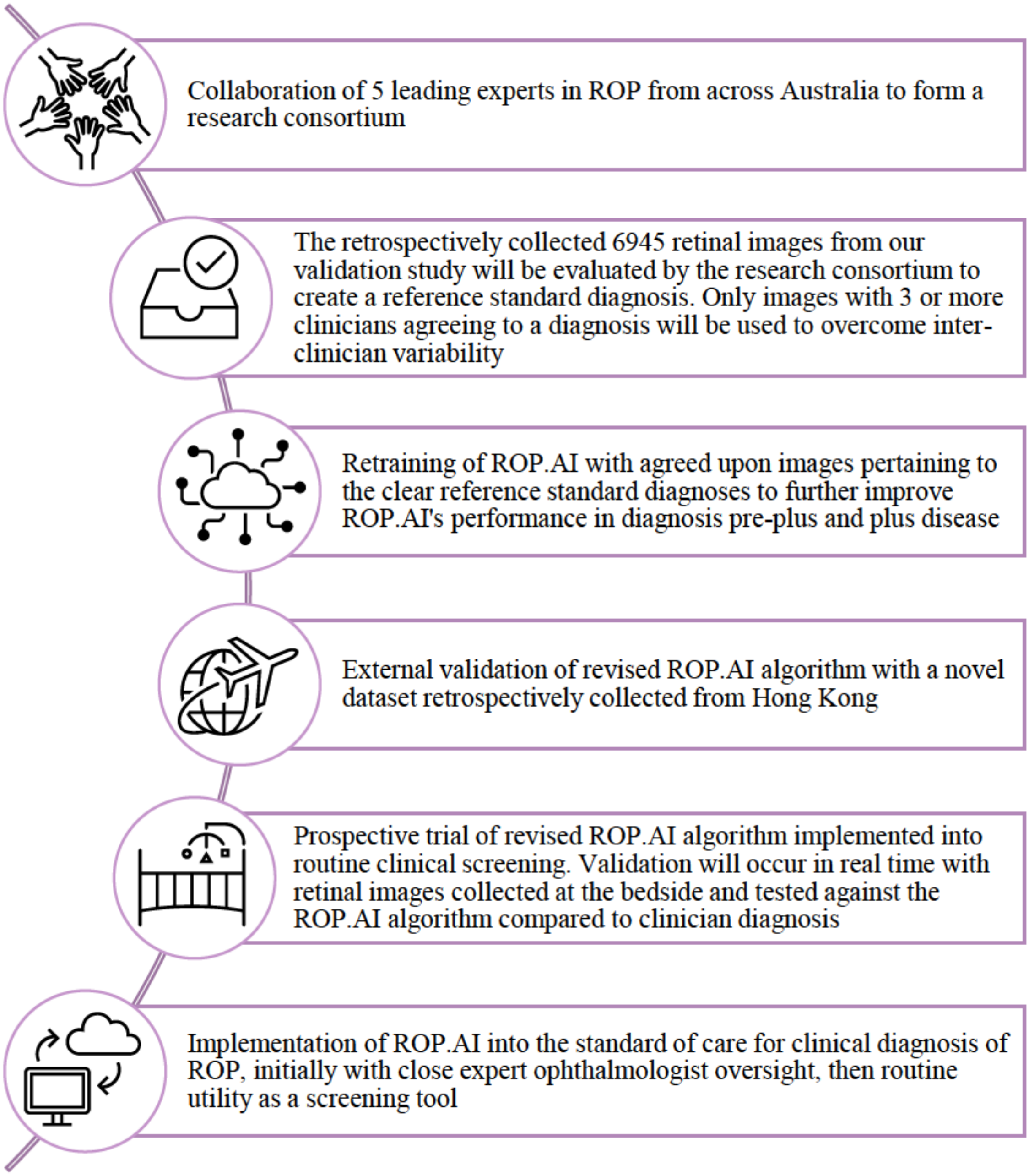


FIGURE 14: Future Directions For ROP.AI Including Plan To Implement Into Routine Clinical Practice

The research consortium, comprising of five expert ophthalmologists from different centres across Australia, will work together to grade all 6945 retinal images collected from our validation study. Each specialist ophthalmologist will grade all images and only images with three or more agreed diagnoses will be used to create a reference standard of diagnosis. These images will then be retrained into the current ROP.AI algorithm with an aim to improve its diagnostic performance. An internal hold-out test set of the reference standard images will be reserved for internal validation, and external validation will occur with retrospective images collected from Hong Kong. Upon successful retraining and validation, the updated ROP.AI algorithm will undergo a prospective clinical trial with implementation into routine clinical screening for use by the bedside. Real-time results and validation will occur as the retinal images collected will be tested against the ROP.AI algorithm and compared to the corresponding clinician diagnosis. Final stages of ROP.AI development will occur with integration of the diagnostic algorithm into standard clinical screening, initially for use alongside clinicians, then ultimately as a stand-alone screening test for the diagnosis of ROP.

Key challenges in the adaptation of ROP.AI into clinical practice

There are several key challenges which contribute to the limited examples of real-world adoption of AI into ROP diagnosis and ophthalmology in general. While promising and robust evidence for deep learning algorithms is increasing in the research domain, there are several caveats to be considered before successful clinical deployment.

Firstly, a key challenge lies in a mindset shift for clinicians to entrust clinical judgement and care to machines. The ‘black box’ approach of deep learning, where image features learned by the algorithm remains unknown to the user, [60] represents a major contributing factor to this. An AI algorithm will learn image features most predictive for classification, rather than explicit detection of the clinical features that clinicians are most familiar with. [60] This lack of transparency remains a major hurdle for both clinician and patient trust, [108-110] especially in clinical support systems where misdiagnosis has significant sequelae. While there are research groups exploring how to make AI more explainable and transparent to the user, [111] there appears to be compromise between accuracy and explainability. Models with clear explanations and high transparency have been shown to perform less accurate than those with the low transparency but high performance. [112] In practice, full transparency may be extremely difficult to achieve. Instead, the focus should be on providing education to clinicians and producing robust, validated results to support AI utilisation alongside clinical care.

ROP.AI may be incorporated into clinical practice in a semi-automated model initially, before entrusting screening completely to AI. For example, retinal images captured by the bed side can be analysed initially by ROP.AI before being flagged as either ‘referable’ or ‘ungradable’, requiring clinician confirmation. The benefit of the ROP disease process is that disease is often scaled on a continuum, [9] and both ROP.AI and other studies have demonstrated AI’s ability to detect this disease spectrum. [69, 73, 103] ROP.AI’s output criteria can therefore be re-categorized or modified into ‘needs clinician input’ based on the degree of its diagnostic agreement between disease and no disease. Processes can also be implemented to deal with any disagreements between the clinician and the algorithm to reduce clinical uncertainty. The successful incorporation of ROP.AI into a semi-automated model may expose its clinical potential and further exploration into a fully-automated model can then occur.

There lie technical challenges which are important to consider in the implementation of an AI screening model. Firstly, automated algorithms will need to be integrated either into commercially available retinal cameras such as RetCam, or into cloud-base systems. This may require collaboration with system developers, engineers, or even further infrastructure to support a telemedicine model. Additionally, all images captured will have patient-related data and needs to be stored and processed in a way that is secure and private. This process will require technical solutions including hardware and software investments that may be time and cost-intensive. Nonetheless, advancements in technology will continue to improve and healthcare and research groups should continue to collaborate to overcome these barriers. There is no doubt that deep learning and AI harness immense power for the advancement and discovery of medical and scientific knowledge and healthcare should not fall behind because to these challenges.

The Future of ROP.AI

ROP.AI holds potential to improve the diagnosis of ROP worldwide, with strict, well-designed research as the backbone to this. ROP diagnosis is known to be both subjective and qualitative, [32] and ROP.AI may improve accuracy and introduce objectivity to disease diagnosis. Additionally, limited workforce is available to meet the demands of screening required in rural and regional areas, and AI based screening may provide the paradigm shift towards improved accessibility. ROP.AI has demonstrated generalizability against a new geographic population with high sensitivity and negative predicative values in detecting plus disease. These results demonstrate promise for future quantitative disease classification, risk stratification, and even post-treatment monitoring for disease recurrence or treatment failure.

The algorithm could be optimised to produce a ‘feature-map’ demonstrating abnormal regions of the retina or vessels to improve both transparency for clinicians, but also to assist clinicians in identifying areas most at risk for disease progression. Deployed into a telehealth model, ROP.AI could transform ROP screening and diagnosis with improved accuracy, objectivity, efficiency, and accessibility. The innovation of deep learning and AI provides an opportunity to make key contributions towards the provision of high quality, sustainable ophthalmic care to patients not just with ROP, but all aspects of Ophthalmology, and clinicians must continue to adapt in this fast paced, revolutionizing field.

Conclusion

ROP remains a sight-threatening condition that is increasing in incidence due to advances in neonatal medicine. Early diagnosis through adequate screening is key to managing this debilitating condition. ROP.AI poses a novel and exciting method of screening for ROP through a deep learning algorithm. With this innovative technology, ROP screening could be made more accessible, more objective, and cheaper. This thesis has outlined the important features required of a generalizable, accurate algorithm and our validation of ROP.AI reveals its potential to be a real-world point of care tool for ROP screening.

APPENDIX

Appendix 1 – Systematic review full search strategy

We show the search strategy for a) Medline OVID b) Pubmed c) Web of Science d) Embase

a) Medline OVID

   "Retinopathy of prematurity" OR "ROP" AND

   "Diagnosis" or "screening" AND

   "Artificial intelligence" OR "deep learning" OR "convolutional neural networks"

b) PubMed

   "Retinopathy of prematurity" OR "ROP" AND "diagnosis" or "screening" AND "artificial intelligence" OR "deep learning" OR "convolutional neural network"

c) Web of Science

   TI = (diagnosis or screening or classification) AND

   TS = (artificial intelligence or machine learning or deep learning or convolutional neural network) AND

   TI = (retinopathy of prematurity or ROP)

d) Embase

   "Retinopathy of Prematurity" OR "ROP" OR "plus disease" AND

   "Diagnosis" or "screening" or "classification" AND

   "Artificial intelligence" OR "deep learning" OR "convolutional neural network" OR "machine learning"

Appendix 2 – Methodological quality assessment of bias for included studies using QUADAS-2 [76]

Study	Domain 1A	Domain 1B	Domain 2A	Domain 2B	Domain 3A	Domain 3B	Domain 4A
Brown et al. 2018	Unclear	Low	Low	Low	Low	Low	Low
Chen et al. 2020	Low	Low	Low	Low	Unclear	Low	Low
Hu et al. 2019	Unclear	Unclear	Low	Low	Low	Low	Low
Huang et al. 2020	Low	Low	Low	Low	Low	Low	Low
Mao et al. 2020	High	Unclear	Low	Low	High	Low	Unclear
Ramachandran et al. 2021	High	Low	Low	Low	Low	Low	Low
Tan et al. 2019	Low	Low	Low	Low	Low	Low	Low
Tong et al.2020	Low	Low	Low	Low	High	High	High
Wang et al. 2021	Low	Low	Low	Low	Unclear	Low	Unclear
Wang et al. 2018	Low	Low	Low	Low	Low	Low	Low
Yildiz et al. 2020	High	Low	Low	High	Low	Low	Low
Zhang et al. 2018	High	Low	Low	Low	Low	Low	Unclear

Appendix 3 – HREC Approval Documents

CHILDREN'S HEALTH QUEENSLAND
HOSPITAL AND HEALTH SERVICE
HUMAN RESEARCH ETHICS COMMITTEE
[redacted]

Queensland Government

Level 7, Centre for Children's Health Research
Queensland Children's Hospital Precinct
62 Graham Street, South Brisbane QLD 4101
[redacted]

1st April 2020

A/Prof Shuan Dai
Director of Opthalmology
Queensland Children's Hospital
501 Stanley Street
South Brisbane QLD 4101

Dear A/Prof Dai,

HREC Reference number: HREC/20/QCHQ/62358
Project title: Real-world Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Retinopathy of Prematurity Project

Many thanks for the submission of the above low risk project and your response to request for further information. This has now been reviewed.

I am pleased to advise the proposal meets the requirements of the National Statement on Ethical Conduct in Human Research (2007) and the Committee is happy to give approval. This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (2007), and all subsequent updates, the Australian Code for the Responsible Conduct of Research (2007) and the CPMP/ICH Note for Guidance on Good Clinical Practice.

This project has Ethics approval for the following sites:

- Queensland Children's Hospital, Brisbane/Centre for Children's Health Research, QLD
- Royal Brisbane and Women's Hospital, QLD
- Mater Children's Private Hospital, QLD
- Perth's Children's Hospital, WA
- Westmead Children's Hospital, NSW
- Royal Children's Hospital Melbourne, VIC

Please note the HREC does not issue pre-determined approval periods. Ethical approval is ongoing, subject to the submission of an annual report on the anniversary of approval.

The documents reviewed and approved include:

Document	*Version*	*Date*
Cover Letter		3 March 2020
Application (62358)		2 March 2020
Protocol	1.0	3 March 2020
Parent / Guardian Consent Form	1.0	2 March 2020
Parent / Guardian information Sheet	1.0	2 March 200

approval 010420

Western Australian Specific Module		27 March 2020
Victorian Specific Module		2 March 2020
Appendix 1- ICROP Classification ROP		
Letter of Support – QCH		4 March 2020

Please note the following conditions of approval:

1. We require an annual progress report (that covers all sites listed on approval) for the duration of the project. This report is due on the anniversary of HREC approval. Continuation of ethics approval is contingent on submission of an annual report. Failure to comply with this requirement may result in suspension of the project by the HREC.
   A comprehensive Final Report upon completion of the project. (In accordance with National Statement 5.5.3)

2. Please note if identifiable or potentially re-identifiable data for this research project is to be accessed without the written consent of the person to whom the data relates an application for disclosure of this data must be made under the *Public Health Act*. Further information regarding the *Public Health Act* is available via this link: **http://www.health.qld.gov.au/ohmr/html/regu/aces_conf_hth_info.asp**

3. Submit to the reviewing HREC for approval any proposed amendments to the project including any proposed changes to the Protocol, Participant Information and Consent Form/s and the Investigator Brochure.

4. If the project does not proceed, the Committee must be informed as soon as possible. (In accordance with National Statement 5.5.6)

5. The Committee must be informed of any potential or realised problem with bioethical implications, if such occurs during the conduct of the research project.

6. Any serious adverse event (SAE) that arises in the context of this research, or involving a researcher conducting this research, must be reported to the Ethics Committee within 72 hours and reported to the sponsor (if applicable) within the stipulated time frame.

   Serious Adverse Event Reports that are generated off-site may be (a) Serious Unexpected Adverse Reactions or (b) Serious Events which the Research Team believes cannot be related to the research intervention. The Research team must report incidents of (a) during multi-centre trials. Such are required to be submitted to the Chair of HREC on receipt by the researcher. A summary of the SAE reports is to accompany the submission. Information required includes; patient details (age & sex), adverse event, outcome and the likelihood of the event being related to the study drug/device/procedure.

   With respect to all SAEs, the researcher must provide his or her opinion as to whether the SAE is directly related to the research intervention. A copy of the SAE Summary must be provided.

8. The HREC, authorising institution and/or their delegate/s may conduct an audit of the project at any time.

9. Ethical approval to undertake this research project is given on the understanding that you have an intention to publish your findings in a refereed journal or similar peer-reviewed forum. If you do not have this intention, it is an absolute requirement that you notify the Ethics Committee formally. In this latter instance, approval for this research is not given at this time; and will require further negotiation. Your work must be in accordance with the following:

   - *National Statement on Ethical Conduct in Human Research:* **https://www.nhmrc.gov.au/guidelines/publications/e72**
   - *Queensland Health Management Research Policy:* **http://www.health.qld.gov.au/ohmr/html/regu/resrch_mge_policy.asp**
   - *Declaration of Helsinki:* **http://www.wma.net/en/30publications/10policies/b3/17c.pdf**
   - *Guidelines under Section 95 of the Privacy Act1995 and Guidelines approved under Section 95A of the Privacy Act 1995.* **http://www.health.qld.gov.au/ohmr/html/regu/aces_conf_hth_info.asp**
   - *Queensland Health Privacy Guidelines IS42 & IS42A:* ***http://www.health.qld.gov.au/privacy/IS42A.asp***

10. Researchers should note, if not QLD Health employees, a Blue Card may be required for contact with children.

11. The Researcher must send the **'Notification of Commencement of Research Protocol'** as soon as research begins. Status of the project will remain as 'Not Started' until this form is received.

approval 010420

12. In accordance with the National Statement (3.3.12), before beginning the clinical phase of the research, researchers should register clinical trials in a publicly accessible domain.

Should you have any queries about the HREC's consideration of your project please contact Mrs Amanda Smith (Co-ordinator) or Professor Alan Isles (Chairperson). The HREC terms of Reference, Standard Operating Procedures, membership and standard forms are available from: http://www.health.qld.gov.au/ohmr/html/regu/regu_home.asp

You are reminded that this letter constitutes ethical approval only.

The HREC wishes you every success in your research.

Yours sincerely,

[redacted]

Professor Alan Isles AM
Chair
Children's Health Queensland
Hospital and Health Service
Human Research Ethics Committee

Cc: Ethics Committee Files

approval 010420

CHILDREN'S HEALTH QUEENSLAND
HOSPITAL AND HEALTH SERVICE
HUMAN RESEARCH ETHICS COMMITTEE
[redacted]

Queensland Government

Level 7, Centre for Children's Health Research
Queensland Children's Hospital Precinct
62 Graham Street, South Brisbane QLD 4101
[redacted]

4th May 2021

Associate Professor Shuan Dai
Director of Ophthalmology
Queensland Children's Hospital
501 Stanley Street
South Brisbane QLD 4101

Dear Professor Dai,

HREC Reference number: HREC/20/QCHQ/62358
Project title: Real-world Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Retinopathy of Prematurity Project

Many thanks for your submission of the 26th April regarding the addition of a site to the above project. This has now been reviewed and I am pleased to advise that the documents reviewed and approved were:

Document	Version	Date
Amendment submission		26 April 2021
Protocol	3.0	16 April 2021
Additional Site: Royal Women's Hospital, Parkville Victoria		
Additional Investigator: Associate Professor James Elder		
Noted update/corrected Site Name: Mater Misercordiae, Brisbane		

Please note that additional Site Specific Assessment Authorisation will be required from this new site before research may commence.

The Children's Health Queensland HREC is constituted and operates in accordance with the National Health and Medical Research Council's *"National Statement on Ethical Conduct in Human Research (2007), NHMRC and Universities Australia Australian Code for the Responsible Conduct of Research (2007)* and the "*CPMP/ICH Note for Guidance on Good Clinical Practice"*.

A copy of this letter must be forwarded to any Research Governance Office that has also approved this project. It should be noted at all requirements of the original approval still apply.

Yours sincerely,

[redacted]

Professor Alan Isles AM
Chair, Children's Health Queensland
Hospital and Health Service
Human Research Ethics Committee

add site 040521

CHILDREN'S HEALTH QUEENSLAND
HOSPITAL AND HEALTH SERVICE
HUMAN RESEARCH ETHICS COMMITTEE

[redacted]

Queensland
Government

Level 7, Centre for Children's Health Research
Queensland Children's Hospital Precinct
62 Graham Street, South Brisbane QLD 4101
[redacted]

4th May 2021

Associate Professor Shuan Dai
Director of Ophthalmology
Queensland Children's Hospital
501 Stanley Street
South Brisbane QLD 4101

Dear Professor Dai,

HREC Reference number: HREC/20/QCHQ/62358
Project title: Real-world Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Retinopathy of Prematurity Project

Many thanks for your submission of the 28th April regarding the addition of a site to the above project. This has now been reviewed and I am pleased to advise that the documents reviewed and approved were:

Document	Version	Date
Amendment submission		28 April 2021
Protocol	3.0	16 April 2021
Additional Site: Gold Coast University Hospital		
Additional Investigator: Dr Heather Russell, Dr Aditi Kirpalani		

Please note that additional Site Specific Assessment Authorisation will be required from this new site before research may commence.

The Children's Health Queensland HREC is constituted and operates in accordance with the National Health and Medical Research Council's *"National Statement on Ethical Conduct in Human Research (2007), NHMRC and Universities Australia Australian Code for the Responsible Conduct of Research (2007)* and the "*CPMP/ICH Note for Guidance on Good Clinical Practice"*.

A copy of this letter must be forwarded to any Research Governance Office that has also approved this project. It should be noted at all requirements of the original approval still apply.

Yours sincerely,

[redacted]

Professor Alan Isles AM
Chair, Children's Health Queensland
Hospital and Health Service
Human Research Ethics Committee

Cc: Ethics Committee Files

add site 2040521

Appendix 4 – SSA Agreements

Queensland Government

Children's Health Queensland
Hospital and Health Service

A/Prof Shuan Dai
Director of Ophthalmology
Queensland Children's Hospital
501 Stanley Street
South Brisbane QLD 4101

Dear Dr Dai,

ERM Project ID:	**62358**
SSA reference number:	**SSA/20/QCHQ/62358**
Project title:	**Real-world Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Retinopathy of Prematurity Project**

Thank you for submitting an application for authorisation of this project. I am pleased to inform you that authorisation has been granted for this study to take place at the following site/s:

CHILDREN'S HEALTH QUEENSLAND HOSPITAL AND HEALTH SERVICE

Please be advised of the allocated Human Research Ethics Committee (HREC) and Site Specific Application reference for your study and quote them when communicating with our office for study identification purposes.

Document	Version	Date
Site Specific Application	Sep Ver 1	16 September 2020

Furthermore, I acknowledge the receipt of the following documents submitted with the SSA application

Document	Version	Date
Research Collaboration Agreement	-	Date of last party signature
PHA approval letter	-	12 October 2020
HREC Approval Letter		1 April 2020 17 September 2020 4 May 2021
Protocol	3.0	16 April 2021
Appendix 1 – ICROP Classification ROP	-	-
Letter of Support – Division of Surgery and Perioperative Services	-	23 June 2020

The following conditions apply to this research proposal. These are additional to those conditions imposed by the HREC that granted ethical approval.

1. The study must be conducted in accordance with the Australian Code for the Responsible Conduct of Research (2008) and ICH Good Clinical Practice (GCP).

Queensland Children's Hospital
PO Box 3474
SOUTH BRISBANE QLD 4101

2. Proposed AMENDMENTS to the research protocol or conduct of the research which may affect the ethical acceptability of the project are to be submitted to the HREC for review. A copy of the HREC approval letter must be submitted to the Research Governance Office via Ethical Review Manager (ERM);
3. Annual reports and final reports are to be submitted to your approving HREC as well as the Research Governance Office on the anniversary of your approval via ERM.
4. Please complete the Notification of Commencement Form once commencement of this protocol has occurred at this site and return to the Research Governance Office and approving HREC via ERM.
5. Please inform the Research Governance Office of any extension to the duration of the project approved by the HREC. Your site authorisation is in conjunction with the validity of your HREC approval.
6. The Research Governance Office may conduct a random audit of your study at any time.

May I take this opportunity to wish you every success with your research project and the contribution it will make to Children's Health Queensland and the broader research community.

Yours Sincerely

[redacted]

Dr Andrew Moore
Director of Research
Children's Health Queensland
Hospital and Health Service

26 / 05 / 2021

Mr Arthur Hui
Manager

the women's
the royal women's hospital

ABN 62 787 822 077
Locked Bag 300

Research and Ethics Secretariat
[redacted]

26.8.21

Associate Professor J Elder
Consultant Paediatric Ophthalmologist
RWH

Dear Professor Elder,

Re: Project 21/38 (SSA 21/19) – Real-world validation of deep learning algorithm ROP AI for the automated diagnosis of Retinopathy of prematurity

HREC Reference Number: HREC/20/QCHQ/62358
EMR Study Code/Identifier: RWH 62358
SSA Project ID: SSA/62358/RWHV-2021-268180 (v1)

Thank you for submitting a Site Specific Assessment Form for authorisation of the above project at The Royal Woman's Hospital.

I am pleased to inform you that authorisation / approval has been granted for this project to be conducted at RWH.

Document	*Version*	*Date*
Protocol	3.0	16.4.21
Human Research Ethics Application		30.3.20
Approval letters from Children's Health Queensland Human Research Ethics Committee		17.9.20 & 4.5.21
Site Specific Assessment		8.6.21

Prior to commencement of your project, you are reminded that you must contact the relevant RWH Divisional Directors / Department Heads to confirm your actual commencement date. Failure to inform these RWH personnel may jeopardise their approval and support for your project.

Please return the completed Notification of Project Commencement Form to me when the project begins.

REMINDER & PLEASE NOTE

If you have not already done so, you are reminded that researchers should complete an EMR (Electronic Medical Record - EPIC) Pre-Submission Survey (for REDCap) for their project to be activated in EMR.

The link researchers should use to complete the EMR Pre-Submission Survey is:
https://biredcap.mh.org.au/surveys/?s=LN7MT4PYF4

This is important. For EMR processes to work, it needs timely notification of research projects and for staff to follow the agreed processes.

1. Use of the Electronic Medical Record (EMR) for research

   Accessing identified patient information from health services other than your own for research purposes is NOT permitted. Precinct polices regarding appropriate access MUST be followed. Identified patient information must ONLY EVER be accessed after governance authorisation or audit/quality assurance (QA) approval from the appropriate health service.

2. Good Clinical Practice (GCP) Training

   The RWH Human Research Ethics Committee has now agreed that GCP training should be mandatory for all principal investigators. The Committee also strongly encourages GCP training for other associate investigators and student researchers.

Yours sincerely,

A. C. B. Hui
Manager
Research and Ethics Secretariat

7 September 2021

Dr Shaheen Shah
Consultant Ophthalmologist
Ophthalmology
Queensland Children's Hospital

Dear Dr Shah

Project Id: 62358
Project Title: Real-world Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Retinopathy of Prematurity
Review Reference: MSSA/MRGO/62358 (V4)

Thank you for submitting an application for authorisation of this project.

I am pleased to inform you that authorisation has been granted for this study to take place at the following Mater Misericordiae Limited site(s):

- Mater Hospital Brisbane
- Mater Mothers' Hospital
- Mater Private Hospital Brisbane

Documents reviewed and authorised by Mater Research Governance are as per those listed on HREC approval letter/s dated 01 April 2020, 17 September 2020 and 04 May 2021.

The following conditions apply to this research project. These are additional to those conditions imposed by the HREC that granted ethical approval.

1. The Mater Research Governance Office must be informed of any problems that arise during the course of the study which may affect conduct of the study at Mater, including serious or unexpected adverse events occurring at the site.
2. Proposed amendments to the research protocol, study documentation, or conduct of the research must be submitted to both the reviewing HREC and the Mater Research Governance Office. Written HREC approval and Mater Research Governance authorisation are both required before an amendment may be implemented at Mater.
3. Proposed amendments to the conduct of the research which affect the ongoing acceptability of the project at Mater (including changes to any of the following: funding and budget, existing research agreements, site investigator team or the Mater contact person) are to be submitted to the Mater Research Governance Office.
4. The Mater Research Governance Office must be notified of any students involved with the study, regardless of their role in the study or the stage of the study at which they are added to the investigator team.
5. The Mater Research Governance Office is responsible for monitoring research authorised for conduct at Mater, and for determining the method of monitoring which is appropriate to each project:

   (a) For all projects: Annual Progress Reports must be provided to the Mater Research Governance Office on the anniversary of the original HREC Approval. Mater Research Governance accepts progress reports on whichever template is required by the reviewing HREC.

   (b) A proportion of all human research projects authorised for conduct at Mater will be eligible for site monitoring each year and the research team will be informed prior to the monitoring visit if this project has been selected.

We wish you every success in undertaking this research.

Yours sincerely

Amanda Sands
Research Governance Officer

Mater Research Governance
Research Compliance | Level 2 | Aubigny Place | South Brisbane | Qld 4101
[redacted] w: materresearch.org.au

Queensland Health

Queensland Government

Enquiries to: **Research Governance Officer**
Phone: [redacted]
Our Ref: ERM ref# 62358 (VD)

Dr Heather Russell
Principal Investigator
Gold Coast University Hospital
[redacted]

HREC reference number: **HREC/2020/QCHQ/62358**
SSA reference number: **SSA/2021/QGC/62358**

Project title: **Real-world Validation of Deep Learning Algorithm ROP.AI for the Automated Diagnosis of Retinopathy of Prematurity Project**

Thank you for submitting an application for authorisation of the above project. I am pleased to inform you that authorisation has been granted for this study to take place at the following site:

- Gold Coast Hospital and Health Service

Ethics approval for this project was granted on 1 April 2020 by the Children's Health Queensland Hospital and Health Service Human Research Ethics Committee

The documents reviewed and authorised that are relevant to this site are listed below:

Document	Version No	Version date
SSA application	1.0	8 June 2021
HREC amendment approval (waiver of consent)	-	17 September 2020
HREC amendment approval (add GCHHS as a site)	-	4 May 2021
Protocol	3.0	16 April 2021
HREC granted waiver As there is no 'confidential information' (defined in s 139 of the Hospital and *Health Board's Act 2011*) being disclosed to any person who is not a 'designated person' (also defined in s139 of that Act) and that each and every disclosure of 'confidential information' to 'designated persons' is for the purpose described in Section 150 of the *Hospital and Health Board's Act 2011*, being evaluating, managing, monitoring or planning 'health services' (defined in s150 of that Act) then you do not have to rely on the Public Health Act approval as an exception to the duty of confidentiality		

The following conditions apply to this research proposal. These are additional to those conditions imposed by the Human Research Ethics Committee that granted ethical approval.

1) Proposed amendments to the research protocol or conduct of the research which may affect the ethical acceptability of the project are to be submitted to the HREC for review. A copy of the HREC approval/rejection letter must be submitted to the RGO;

Office
Office for Research Governance & Development
Level 2, Pathology and Education Building
1 Hospital Boulevard
Southport QLD 4215

Phone
[redacted]

2) **Proposed amendments to the research protocol or conduct of the research which only affects the ongoing site acceptability of the project, are to be submitted to the research governance officer;**
3) **Proposed amendments to the research protocol or conduct of the research which may affect both the going ethical acceptability of the project and the site acceptability of the project are to be submitted firstly to the HREC for review and then to the research governance officer after a HREC decision is made.**

If any matters arise concerning the conduct of the research at this site, please ensure you contact the Research Governance Officer at: GCHResearch@health.qld.gov.au **or on 07 5687 8311.**

Yours sincerely

[redacted]

Dr Trevor Fayers
Clinical Director
Surgical, Anaesthetics and Procedural Services
Gold Coast Hospital and Health Service
30-Jun-2021 | 8:12 AM AEST

Office
Office for Research Governance & Development
Level 2, Pathology and Education Building
1 Hospital Boulevard
Southport QLD 4215

Phone
[redacted]

Queensland Government

Metro North
Hospital and Health Service

A/Prof Shuan Dai
Ophthalmology Department
Queensland Children's Hospital
501 Stanley Street
SOUTH BRISBANE QLD 4101
[redacted]

Dear A/Prof Dai,

HREC reference number: HREC/2020/QCHQ/62358
SSA reference number: SSA/2021/QRBW/62358
Project title: Real-world Validation of Deep Learning Algorithm ROP. AI for the Automated Diagnosis of Retinopathy of Prematurity

Thank you for submitting an application for authorisation of this project. I am pleased to inform you that authorisation has been granted for this study to take place at the following site/s:

Royal Brisbane and Women's Hospital

The following conditions apply to this research project. These are additional to those conditions imposed by the Human Research Ethics Committee (HREC) that granted ethical approval.

1. The study must be conducted in accordance with the Australian Code for the Responsible Conduct of Research (2007) and ICH Good Clinical Practice (GCP).
2. Amendments to the research are to be submitted to the HREC for review. A copy of the HREC approval/rejection letter together with the documents must be submitted to the Research Governance Officer (RGO) at each site for acknowledgment/approval.
3. An annual report acknowledged by the HREC must be submitted to the RGO for site records.
4. A final report acknowledged by the HREC must be submitted to the RGO on completion or early termination of the study.
5. All local Serious Adverse Events (SAE) or Suspected Unexpected Serious Adverse Reaction (SUSAR) must be reported to the relevant RGO. Furthermore, the relevant safety reporting requirements of the approving HREC must be adhered to.
6. Informed consent or consent to contact must be initially obtained from a Metro North Hospital and Health Service employee, and not external personnel, unless another authorised provision is applicable.

The documents reviewed and approved include:

Document	*Version*	*Date*
Site Specific Assessment Form	Apr Ver 2	28 April 2021
PHA		12 October 2020

MNHHS Research Services
Level 12 Block 7, RBWH Campus
HERSTON Q 4029

[redacted]

Furthermore, I acknowledge receipt of the following documents submitted with the SSA application:

Document	*Version*	*Date*
HREC Approval Letter		1 April 2020
Application		2 March 2020
Appendix 1 – ICROP Classification ROP	-	-
HREC Amendment Letter		17 September 2020
Protocol	2	16 September 2020

Please complete the Notification of Commencement Form once commencement of this protocol has occurred at this site and return to the RGO and the approving HREC.

I wish you every success in your research project.

Yours sincerely

[redacted]

Professor Michael O'Sullivan
Executive Director Research
Royal Brisbane and Women's Hospital
METRO NORTH HOSPITAL AND HEALTH SERVICE
Dated 31/5/ 2021

REFERENCES

References

1. Blencowe, H., J.E. Lawn, T. Vazquez, et al., *Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010.* Pediatr Res, 2013. **74 Suppl 1**(Suppl 1): p. 35-49.
2. Dai, S., N. Austin, and B. Darlow, *Retinopathy of prematurity: New Zealand recommendations for case detection and treatment.* J Paediatr Child Health, 2015. **51**(10): p. 955-9.
3. 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): p. 77-82.
4. Fleck, B.W. and N. McIntosh, *Pathogenesis of retinopathy of prematurity and possible preventive strategies.* Early Hum Dev, 2008. **84**(2): p. 83-8.
5. Hughes, S., H. Yang, and T. Chan-Ling, *Vascularization of the human fetal retina: roles of vasculogenesis and angiogenesis.* Invest Ophthalmol Vis Sci, 2000. **41**(5): p. 1217-28.
6. Rivera, J.C., P. Sapieha, J.S. Joyal, et al., *Understanding retinopathy of prematurity: update on pathogenesis.* Neonatology, 2011. **100**(4): p. 343-53.
7. Smith, L.E., *Pathogenesis of retinopathy of prematurity.* Growth Horm IGF Res, 2004. **14 Suppl A**: p. S140-4.
8. Pérez-Muñuzuri, A., J.R. Fernández-Lorenzo, M.L. Couce-Pico, et al., *Serum levels of IGF1 are a useful predictor of retinopathy of prematurity.* Acta Paediatr, 2010. **99**(4): p. 519-25.
9. Chiang, M.F., G.E. Quinn, A.R. Fielder, et al., *International Classification of Retinopathy of Prematurity, Third Edition.* Ophthalmology, 2021.
10. Castillo, A., A. Deulofeut R Fau - Critz, A. Critz A Fau - Sola, et al., *Prevention of retinopathy of prematurity in preterm infants through changes in clinical practice and SpO_2technology.* 2010(1651-2227 (Electronic)).
11. Tin, W., P. Milligan Dw Fau - Pennefather, E. Pennefather P Fau - Hey, et al., *Pulse oximetry, severe retinopathy, and outcome at one year in babies of less than 28 weeks gestation.* 2001(1359-2998 (Print)).
12. The BOOSTII UK, A., and NZ Collaborative Groups. , W.O. Stenson Bj Fau - Tarnow-Mordi, B.A. Tarnow-Mordi Wo Fau - Darlow, et al., *Oxygen saturation and outcomes in preterm infants.* 2013(1533-4406 (Electronic)).
13. Carlo, W.A., N.N. Finer, M.C. Walsh, et al., *Target ranges of oxygen saturation in extremely preterm infants.* N Engl J Med, 2010. **362**(21): p. 1959-69.
14. Schmidt, B., R.K. Whyte, E.V. Asztalos, et al., *Effects of targeting higher vs lower arterial oxygen saturations on death or disability in extremely preterm infants: a randomized clinical trial.* Jama, 2013. **309**(20): p. 2111-20.
15. Darlow, B.A., M. Vento, M. Beltempo, et al., *Variations in Oxygen Saturation Targeting, and Retinopathy of Prematurity Screening and Treatment Criteria in Neonatal Intensive Care Units: An International Survey.* Neonatology, 2018. **114**(4): p. 323-331.

16. Good, W.V., *Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial.* Trans Am Ophthalmol Soc, 2004. **102**: p. 233-48; discussion 248-50.
17. Tasman, W., *Multicenter trial of cryotherapy for retinopathy of prematurity.* Arch Ophthalmol, 1988. **106**(4): p. 463-4.
18. Shah, P.K., V. Prabhu, S.S. Karandikar, et al., *Retinopathy of prematurity: Past, present and future.* World J Clin Pediatr, 2016. **5**(1): p. 35-46.
19. Kusaka, S., C. Shima, K. Wada, et al., *Efficacy of intravitreal injection of bevacizumab for severe retinopathy of prematurity: a pilot study.* Br J Ophthalmol, 2008. **92**(11): p. 1450-5.
20. Mintz-Hittner, H.A., K.A. Kennedy, and A.Z. Chuang, *Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity.* N Engl J Med, 2011. **364**(7): p. 603-15.
21. Hunter, D.G. and M.X. Repka, *Diode laser photocoagulation for threshold retinopathy of prematurity. A randomized study.* Ophthalmology, 1993. **100**(2): p. 238-44.
22. McNamara, J.A., W. Tasman, G.C. Brown, et al., *Laser photocoagulation for stage 3+ retinopathy of prematurity.* Ophthalmology, 1991. **98**(5): p. 576-80.
23. Fierson, W.M., *Screening Examination of Premature Infants for Retinopathy of Prematurity.* Pediatrics, 2018. **142**(6).
24. Tan, Z., C. Chong, B. Darlow, et al., *Visual impairment due to retinopathy of prematurity (ROP) in New Zealand: a 22-year review.* Br J Ophthalmol, 2015. **99**(6): p. 801-6.
25. Gilbert, C., A. Fielder, L. Gordillo, et al., *Characteristics of infants with severe retinopathy of prematurity in countries with low, moderate, and high levels of development: implications for screening programs.* Pediatrics, 2005. **115**(5): p. e518-25.
26. Quinn, G.E., *Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic.* Eye Brain, 2016. **8**: p. 31-36.
27. National Neonatology Forum, I., *Evidence Based Clinical Practice Guidelines* 2010.
28. Council, N.H.a.M.R., *Care Around Preterm Birth*, in *Clinical Practice Guideline*. 1997, Commonwealth of Australia.
29. Health, Q., *Consensus statement for screening of retinopathy of prematurity in Queensland.* 2021.
30. Service, G.o.W.A.C.a.A.H., *Retinopathy of Prematurity (ROP) Screening, Treatment and Ophthalmology Consultations Guideline*, in *Neonatology*.
31. Joseph Tauro, S.L.H.D., *Royal Prince Alfred Hospital Guideline: Retinopathy of Prematurity*, N. Government, Editor. 2018.
32. Gschließer, A., E. Stifter, T. Neumayer, et al., *Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity.* Am J Ophthalmol, 2015. **160**(3): p. 553-560.e3.
33. Campbell, J.P., E. Ataer-Cansizoglu, V. Bolon-Canedo, et al., *Expert Diagnosis of Plus Disease in Retinopathy of Prematurity From Computer-Based Image Analysis.* JAMA Ophthalmol, 2016. **134**(6): p. 651-7.
34. Ghergherehchi, L., S.J. Kim, J.P. Campbell, et al., *Plus Disease in Retinopathy of Prematurity: More Than Meets the ICROP?* Asia-Pacific Journal of Ophthalmology, 2018. **7**(3): p. 152-155.
35. Fleck, B.W., C. Williams, E. Juszczak, 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): p. 74-80.

36. Wallace, D.K., G.E. Quinn, S.F. Freedman, et al., *Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity.* J aapos, 2008. **12**(4): p. 352-6.
37. Dai, S., K. Chow, and A. Vincent, *Efficacy of wide-field digital retinal imaging for retinopathy of prematurity screening.* Clin Exp Ophthalmol, 2011. **39**(1): p. 23-9.
38. Welfare, A.I.o.H.a., *Australia's Mothers and Babies 2017-in Brief.*, in *Perinatal statistics series no. 35. Cat. no. PER 100.* . 2019: Canberra: AIHW.
39. Chow, S.S.W., P. Creighton, G.M. Chambers, et al., *Report of the Australian and New Zealand Neonatal Network 2017.* . 2019.
40. Health, D.o., *Australia's Future Health Workforce - Ophthalmology*. 2018.
41. Statistics, A.B.o., *Australian Statistical Geography Standard (ASGS) - Remoteness Structure.* 1270.055.005, 2013. **5**.
42. Yu, T.Y., T. Donovan, N. Armfield, et al., *Retinopathy of prematurity: the high cost of screening regional and remote infants.* Clin Exp Ophthalmol, 2018. **46**(6): p. 645-651.
43. Isaac, M., W. Isaranuwatchai, and N. Tehrani, *Cost analysis of remote telemedicine screening for retinopathy of prematurity.* Can J Ophthalmol, 2018. **53**(2): p. 162-167.
44. Rothschild, M.I., R. Russ, K.A. Brennan, et al., *The Economic Model of Retinopathy of Prematurity (EcROP) Screening and Treatment: Mexico and the United States.* Am J Ophthalmol, 2016. **168**: p. 110-121.
45. Mills, M.D., *Retinopathy of prematurity malpractice claims.* Arch Ophthalmol, 2009. **127**(6): p. 803-4.
46. Altersitz, K. and M. Piechocki, *Survey: Physicians being driven away from ROP treatment*. 2006.
47. Simkin, S.K., S.L. Misra, J.V. Han, et al., *Auckland regional telemedicine retinopathy of prematurity screening network: A 10-year review.* Clin Exp Ophthalmol, 2019. **47**(9): p. 1122-1130.
48. Fierson, W.M. and A. Capone, Jr., *Telemedicine for evaluation of retinopathy of prematurity.* Pediatrics, 2015. **135**(1): p. e238-54.
49. Choi, R.Y., J.M. Brown, J. Kalpathy-Cramer, et al., *Real world variability in plus disease identified using a deep learning-based retinopathy of prematurity severity scale.* Ophthalmol Retina, 2020.
50. Antaki, F., K. Bachour, T.N. Kim, et al., *The Role of Telemedicine to Alleviate an Increasingly Burdened Healthcare System: Retinopathy of Prematurity.* Ophthalmol Ther, 2020.
51. Brady, C.J., S. D'Amico, and J.P. Campbell, *Telemedicine for Retinopathy of Prematurity.* Telemed J E Health, 2020. **26**(4): p. 556-564.
52. Turing, A., *Computing Machinery and Intelligence.* Mind, 1950. **49**: p. 433-460.
53. J McCarthy, M.M., N Rochester, C Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* 1955.
54. Wu, J., E. Yılmaz, M. Zhang, et al., *Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition.* Front Neurosci, 2020. **14**: p. 199.
55. Popel, M., M. Tomkova, J. Tomek, et al., *Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals.* Nat Commun, 2020. **11**(1): p. 4381.
56. Fayyad, J., M.A. Jaradat, D. Gruyer, et al., *Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review.* Sensors (Basel), 2020. **20**(15).

57. Wang, X., Y. Peng, L. Lu, et al., *ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,* C. University, Editor. 2017.
58. Kulkarni, S., N. Seneviratne, M.S. Baig, et al., *Artificial Intelligence in Medicine: Where Are We Now?* Acad Radiol, 2020. **27**(1): p. 62-70.
59. Thian, Y.L., Y. Li, P. Jagmohan, et al., *Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs.* Radiol Artif Intell, 2019. **1**(1): p. e180001.
60. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
61. Ramachandran, N., S.C. Hong, M.J. Sime, et al., *Diabetic retinopathy screening using deep neural network.* Clin Exp Ophthalmol, 2018. **46**(4): p. 412-416.
62. Gulshan, V., L. Peng, M. Coram, et al., *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.* Jama, 2016. **316**(22): p. 2402-2410.
63. Ting, D.S.W., C.Y. Cheung, G. Lim, et al., *Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes.* Jama, 2017. **318**(22): p. 2211-2223.
64. Zhang, Y., J. Shi, Y. Peng, et al., *Artificial intelligence-enabled screening for diabetic retinopathy: a real-world, multicenter and prospective study.* BMJ Open Diabetes Res Care, 2020. **8**(1).
65. De Fauw, J., J.R. Ledsam, B. Romera-Paredes, et al., *Clinically applicable deep learning for diagnosis and referral in retinal disease.* Nat Med, 2018. **24**(9): p. 1342-1350.
66. He, J., T. Cao, F. Xu, et al., *Artificial intelligence-based screening for diabetic retinopathy at community hospital.* Eye (Lond), 2020. **34**(3): p. 572-576.
67. Ming, S., K. Xie, X. Lei, et al., *Evaluation of a novel artificial intelligence-based screening system for diabetic retinopathy in community of China: a real-world study.* Int Ophthalmol, 2021. **41**(4): p. 1291-1299.
68. Brown, J.M., J.P. Campbell, A. Beers, et al., *Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks.* JAMA Ophthalmology, 2018. **136**(7): p. 803-810.
69. Wang, J., R. Ju, Y. Chen, et al., *Automated retinopathy of prematurity screening using deep neural networks.* EBioMedicine, 2018. **35**: p. 361-368.
70. Prematurity., I.C.f.t.C.o.R.o., *The International Classification of Retinopathy of Prematurity revisited.* Arch Ophthalmol, 2005. **123**(7): p. 991-9.
71. Burlina, P.M., N. Joshi, M. Pekala, et al., *Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks.* JAMA Ophthalmol, 2017. **135**(11): p. 1170-1176.
72. Li, Z., Y. He, S. Keel, et al., *Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs.* Ophthalmology, 2018. **125**(8): p. 1199-1206.
73. Tan, Z., S. Simkin, C. Lai, et al., *Deep Learning Algorithm for Automated Diagnosis of Retinopathy of Prematurity Plus Disease.* Transl Vis Sci Technol, 2019. **8**(6): p. 23.
74. Valentine, P.H., J.C. Jackson, R.E. Kalina, et al., *Increased survival of low birth weight infants: impact on the incidence of retinopathy of prematurity.* Pediatrics, 1989. **84**(3): p. 442-5.

75. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L. , *ImageNet Large Scale Visual Recognition Challenge.* International Journal of Computer Vision, 2015. **115**: p. 211-252.
76. Whiting, P.F., A.W. Rutjes, M.E. Westwood, et al., *QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.* Ann Intern Med, 2011. **155**(8): p. 529-36.
77. Page, M.J., J.E. McKenzie, P.M. Bossuyt, et al., *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews.* Int J Surg, 2021. **88**: p. 105906.
78. Moons, K.G., J.A. de Groot, W. Bouwmeester, et al., *Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist.* PLoS Med, 2014. **11**(10): p. e1001744.
79. Chen, J.S., A.S. Coyner, S. Ostmo, et al., *Deep Learning for the Diagnosis of Stage in Retinopathy of Prematurity: Accuracy and Generalizability across Populations and Cameras.* Ophthalmol Retina, 2021.
80. Campbell, J.P., P. Singh, T.K. Redd, et al., *Applications of Artificial Intelligence for Retinopathy of Prematurity Screening.* Pediatrics, 2021. **147**(3).
81. Hu, J., Y. Chen, J. Zhong, et al., *Automated Analysis for Retinopathy of Prematurity by Deep Neural Networks.* IEEE Transactions on Medical Imaging, 2019. **38**(1): p. 269-279.
82. Huang, Y.P., H. Basanta, E.Y. Kang, et al., *Automated detection of early-stage ROP using a deep convolutional neural network.* Br J Ophthalmol, 2020.
83. Mao, J., Y. Luo, L. Liu, et al., *Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks.* Acta Ophthalmol, 2020. **98**(3): p. e339-e345.
84. Ramachandran, S., P. Niyas, A. Vinekar, et al., *A deep learning framework for the detection of Plus disease in retinal fundus images of preterm infants.* Biocybernetics and Biomedical Engineering, 2021. **41**(2): p. 362-375.
85. Tong, Y., W. Lu, Q.Q. Deng, et al., *Automated identification of retinopathy of prematurity by image-based deep learning.* Eye Vis (Lond), 2020. **7**: p. 40.
86. Wang, J., J. Ji, M. Zhang, et al., *Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening.* JAMA Netw Open, 2021. **4**(5): p. e218758.
87. Yildiz, V.M., P. Tian, I. Yildiz, 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): p. 10.
88. Zhang, Y.S., L. Wang, Z.Q. Wu, et al., *Development of an Automated Screening System for Retinopathy of Prematurity Using a Deep Neural Network for Wide-Angle Retinal Images.* Ieee Access, 2019. **7**: p. 10232-10241.
89. Coyner, A.S., R. Swan, J.P. Campbell, et al., *Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks.* Ophthalmology Retina, 2019. **3**(5): p. 444-450.
90. Greenwald, M.F., I.D. Danford, M. Shahrawat, et al., *Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity.* J aapos, 2020.
91. Gupta, K., J.P. Campbell, S. Taylor, et al., *A Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning to Monitor Disease Regression After Treatment.* JAMA Ophthalmol, 2019. **137**(9): p. 1029-36.

92. Redd, T., J.P. Campbell, J.M. Brown, et al., *Utilization of a deep learning image assessment tool for epidemiologic surveillance of retinopathy of prematurity.* Investigative Ophthalmology and Visual Science, 2019. **60**(9).
93. Smith, K.N., S.J. Kim, I. Goldstein, et al., *Quantitative analysis of aggressive posterior retinopathy of prematurity using deep learning.* Investigative Ophthalmology and Visual Science, 2019. **60**(9).
94. Taylor, S., K. Gupta, J.P. Campbell, et al., *Automated computer-based image analysis in monitoring disease progression for retinopathy of prematurity.* Investigative Ophthalmology and Visual Science, 2018. **59**(9).
95. Wallace, D.K., Z. Zhao, and S.F. Freedman, *A pilot study using "ROPtool" to quantify plus disease in retinopathy of prematurity.* J aapos, 2007. **11**(4): p. 381-7.
96. Wang, J., G. Zhang, J. Lin, et al., *Application of standardized manual labeling on identification of retinopathy of prematurity images in deep learning.* Zhonghua Shiyan Yanke Zazhi/Chinese Journal of Experimental Ophthalmology, 2019. **37**(8): p. 653-657.
97. Campbell, J.P., R.V.P. Chan, S. Ostmo, et al., *Analysis of the relationship between retinopathy of prematurity zone, stage, extent and a deep learning-based vascular severity scale.* Investigative Ophthalmology and Visual Science, 2020. **61**(7).
98. Ramachandran, S., S. Kochitty, A. Vinekar, et al., *A fully convolutional neural network approach for the localization of optic disc in retinopathy of prematurity diagnosis.* Journal of Intelligent & Fuzzy Systems, 2020. **38**(5): p. 6269-6278.
99. Worrall DE, W.C., Brostow GJ, *Automated retinopathy of prematurity case detection with convolutional neural networks.* In: Deep learning and data labeling for medical applications, 2016: p. 68-76.
100. Ataer-Cansizoglu, E., V. Bolon-Canedo, J.P. Campbell, et al., *Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the "i-ROP" System and Image Features Associated With Expert Diagnosis.* Translational Vision Science & Technology, 2015. **4**(6).
101. Touch, P., Y. Wu, Y. Kihara, et al., *Development of AI deep learning algorithms for the quantification of retinopathy of prematurity.* Journal of Investigative Medicine, 2019. **67**(1): p. 209.
102. Wang, S., Y. Zhang, S. Lei, et al., *Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: a systematic review and meta-analysis of diagnostic test accuracy.* Eur J Endocrinol, 2020. **183**(1): p. 41-49.
103. Redd, T.K., J.P. Campbell, J.M. Brown, et al., *Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity.* Br J Ophthalmol, 2018.
104. Chen, J., J.P. Campbell, S. Ostmo, et al., *Automated assessment of stage in retinopathy of prematurity using deep learning.* Investigative Ophthalmology and Visual Science, 2020. **61**(7).
105. Brown, J.M., J.P. Campbell, A. Beers, et al., *Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks.* JAMA Ophthalmol, 2018. **136**(7): p. 803-810.
106. Bolón-Canedo, V., E. Ataer-Cansizoglu, D. Erdogmus, et al., *Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach.* Computer Methods and Programs in Biomedicine, 2015. **122**(1): p. 1-15.
107. Erdogmus, D., V. Bolon-Canedo, E. Ataer-Cansizoglu, et al., *Dealing with inter-expert variability in retinopathy of prematurity through machine learning.* Investigative Ophthalmology and Visual Science, 2015. **56**(7): p. 5251.

108. Elshawi, R., M.H. Al-Mallah, and S. Sakr, *On the interpretability of machine learning-based model for predicting hypertension.* BMC Med Inform Decis Mak, 2019. **19**(1): p. 146.
109. C.Gretton, *Trust and Transparency in Machine Learning-Based Clinical Decision Support.* Human and Machine Learning 2018: p. 279-292.
110. Li, J.O., H. Liu, D.S.J. Ting, et al., *Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective.* Progress in Retinal & Eye Research, 2021. **82**: p. 100900.
111. Adadi, A. and M. Berrada, *Peeking inside the black- box: A survey on Explainable Artificial Intelligence (XAI).* IEEE, 2018. **6**: p. 52138-52160.
112. Holzinger, A., C. Biemann, C. Pattichis, et al., *What do we need to build explainable AI systems for the medical domain?* Explainable AI for the Medical Domain, 2017.