Accepted: 11 August 2020

DOI: 10.1111/dme.14386

DIABETIC Medicine

RESEARCH: CARE DELIVERY

THEIA™ development, and testing of artificial intelligence-based primary triage of diabetic retinopathy screening images in New Zealand

E. Vaghefi[1,2] | **S. Yang**[2,3] | **L. Xie**[2] | **S. Hill**[4] | **O. Schmiedel**[5] | **R. Murphy**[6]* | **D. Squirrell**[1,4]*

[1]Toku Eyes, Auckland, New Zealand

[2]School of Optometry and Vision Science, Auckland, New Zealand

[3]School of Computer Sciences, University of Auckland, Auckland, New Zealand

[4]Department of Ophthalmology, Aukland, New Zealand

[5]Auckland Diabetes Centre, Auckland District Health Board

[6]School of Medicine, Auckland, New Zealand

Correspondence
Ehsan Vaghefi, Toku Eyes, Auckland, New Zealand.
Email: e.vaghefi@auckland.ac.nz

Funding information
This work was funded by the Ministry of Business, Innovation and Education of New Zealand (UOAX1805 - 3715780).

Abstract

Aim: To develop and evaluate an artificial intelligence triage system with high sensitivity for detecting referable diabetic retinopathy and maculopathy, while maintaining high specificity for non-referable disease, for clinical implementation within the New Zealand national diabetic retinopathy screening programme.

Methods: The THEIA™ artificial intelligence system for retinopathy and maculopathy screening, was developed at Toku Eyes using routinely collected retinal screening datasets from two of the largest district health boards in Auckland, New Zealand: the Auckland District Health Board and the Counties Manukau District Health Board. All retinal images from consecutive individuals receiving retinal screening between January 2009 and December 2018 were used. Images were labelled as non-sight-threatening, potentially referable or sight-threatening for New Zealand implementation, or as referable (potentially referable + sight-threatening)/non-referable (non-sight-threatening) for global comparison.

Results: Data from 32 354 unique people with diabetes (63 843 when including multiple visits) were available, of which 95–97%, 0.9–2.4% and 1.1–3.1% were categorized as non-sight-threatening, potentially referable and sight-threatening, respectively. Using the referable/non-referable categories, THEIA achieved overall sensitivity of 94% (95% CI 92–95) in the Auckland District Health Board and 95% (95% CI 92–97) in the Counties Manukau District Health Board datasets, while preserving specificity of 63% (95% CI 62–64) for the Auckland District Health Board and 61% (95% CI 60–62) for the Counties Manukau District Health Board. Implementing THEIA into a New Zealand national diabetic screening programme could significantly reduce the manual grading load.

Conclusion: THEIA, an artificial intelligence tool to assist in clinical decision-making, tailored to the needs of the New Zealand national diabetic screening programme, delivered high sensitivity for detecting referable retinopathy within the multi-ethnic New Zealand population with diabetes.

R.M. and D.S. are equally contributing senior authors.

1 | INTRODUCTION

Artificial intelligence (AI) has progressed rapidly during the past decade with the advent of deep learning. The field of ophthalmology has been an early adopter of these technologies[1–4] and possibly the most promising application of AI in ophthalmology is as a screening tool for detecting diabetic retinopathy (DR). The accuracy of AI-based models for detecting DR has been demonstrated in many previous studies.[3] The results from a landmark prospective study evaluating the performance of a DR diagnostic system in a primary care setting represented an important clinical milestone, as these results were used to form the basis of the first fully autonomous AI-based system approved by the US Food and Drug Administration.[5] However, there are still many clinical and technical challenges with regard to clinical implementation. Firstly, there is a problem with generalizability as this and many other studies have used training datasets from relatively homogenous populations. Secondly, many studies focus on training algorithms to simply distinguish between non-referable and referable DR and do not distinguish between retinopathy and maculopathy. Thirdly, there is a concern about the 'black box' phenomenon of many AI systems, a term which refers to the inability of the user to determine how the AI derived its output. Arguably, it is difficult to ask people with diabetes, clinicians and regulators to trust a system when its workings and thus its inherent biases are unknown.[3,6] Finally, besides the technical challenges, there are also a number of legal and ethical issues that need to be addressed before AI is implemented within healthcare environments.

In addressing these issues, we have firstly combined historic data of the two largest New Zealand (NZ) district health boards, covering 22% of the country's population and ensuring that our AI (named THEIA) was trained and tested on data that are representative of the NZ general population.[7] Secondly, we have created THEIA to be compatible with NZ Ministry of Health standards for diabetic eye screening (Table S1), and, thirdly, we have designed THEIA so that it generates 'attention maps' for its grading decisions, allowing clinicians to examine the basis of the AI-generated grades.

With these challenges in mind, we designed THEIA as a primary triage tool, in effect, allowing the New Zealand screening programme to transition to a semi-automated model of care, with THEIA being used to safely and rapidly triage two groups: (1) people with diabetes with none or minimal disease who could be issued their results at the time of screening and (2) those with sight-threatening disease whose images needed urgent review by the human grading team. This then left a third group of images, comprising people who had mild to moderate disease which may require onward referral but did not require urgent review by the secondary human grading team (Figure 1). Thus, the anticipated position of THEIA, relative to the current DR screening pathway, was as an initial triage tool, designed to reduce the number of images being reviewed manually.

What's new?

- We have developed and validated an artificial intelligence (AI) system to detect referable diabetic retinopathy and maculopathy, using independent screening datasets covering 25% of the New Zealand population.
- We have evaluated the clinical load-saving capacity of this AI system within the New Zealand national diabetic retinopathy screening programme.
- This system provides an automated decision rule to ensure rapid, accurate classification of the large proportion of normal images from the few with abnormal features for prompt, accurate clinical grading, but is not designed to replicate a screening programme.

Specifically, our aims were to: (1) train AI (THEIA) to detect referable DR and maculopathy then validate it using two independent DR screening datasets—the Auckland District Health Board (ADHB) and Counties Manukau District Health Board (CMDHB) dataset—each with different cameras, disease profiles and patient demographics, and (2) to determine the diagnostic performance of THEIA for the automated detection of non-referable, and referable diabetic eye disease.

2 | METHODS

2.1 | Study population

This was a retrospective study using all consecutive retinal screening images acquired as part of routine diabetes photoscreening between January 2009 and December 2018, from two district health boards within the Auckland region: ADHB and CMDHB. These two organizations provide public-funded services to approximately 1 108 850 people (2018/2019 data), which represents approximately 68% of Auckland's population or 22% of the entire New Zealand population.

2.2 | Ethics

The study protocol was approved by the Health and Disability Ethics Committee at New Zealand Health, and the Disability Ethics Committees at the ADHB (Eye-AI 18/CEN/124 and Eye-AI A+8335) and at the CMDHB (Eye-AI

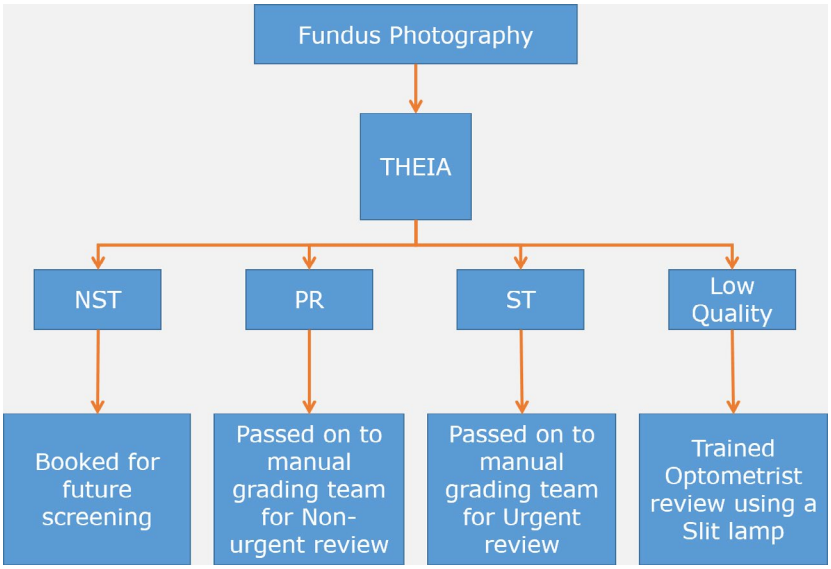

FIGURE 1 Flow chart of outputs generated by THEIA. PR, potentially referable; NST, non-sight-threatening; ST, sight-threatening

947). Appropriate protocols were embedded within THEIA to comply with both Māori data sovereignty and local data privacy regulations.[8–11]

2.3 | Study design, sample size and power

To address any class imbalance issue of our dataset, the weighted loss function strategy was adopted.[12,13] An initial 5000 images, 500 images from each of the five DR severity categories (according to NZ grading standards[14]) and 500 images from each of the five maculopathy severity categories, were randomly selected for AI training. These 5000 images were then re-graded by a retinal specialist (D.S.) and, where a discrepancy arose between the original grade, the image was sent for adjudication before a 'final' training grade was issued. These data were then split in a 70%, 15% and 15% ratio for training, validation and testing, respectively. THEIA's performance was then evaluated on both the ADHB and CMDHB datasets to determine its accuracy, specificity and sensitivity, and generalizability. In NZ, every citizen is assigned a unique National Health Index number. The training and test datasets were separated at the patient level, using this index to avoid data leakage between sets. There were no duplicate images in the datasets used for training or testing.

2.4 | Data capture and reference standard retinal image grading/classification

The NZ Ministry of Health standard mandates that two fundus images per eye be acquired from people with type 2 diabetes (macular and disc-centred) and four fundus images are taken per eye for people with type 1 diabetes (as above, plus inferior and superior to the optic disc). Neither of the two validation datasets from the ADHB or CMDHB were manually curated before being presented to THEIA. Both datasets were collected independently and the images were obtained from multiple different models of fundus camera.

All images were graded according to the NZ Ministry of Health standard[14] by primary and secondary grading teams at their respective district health board, and audited by the lead ophthalmologist of the respective screening programmes. Based on the outcome of screening, the patient is either directly re-enrolled into screening, sent to the tertiary grader for adjudication, or referred to the eye clinic directly (Table 1).

TABLE 1 New Zealand diabetic eye screening standard for primary grading and patient outcome[14]

	Retinopathy	Maculopathy	Patient outcome
Healthy/Non-sight-threatening	R0, R1, R2	M0, M1, M2	Re-enrolled into screening
Potentially referable	R3	M3	Sent for adjudication
Referable (Sight-threatening)	R4, R5	M4, M5	Referred to the eye clinic

2.5 | Image processing and development of the THEIA algorithm

The fundus images were first cropped and resized to 800×800 pixel size. These were then enhanced by using a Gaussian blur technique[15] before being passed into the THEIA algorithm. All image pre-processing steps were fully automated. THEIA comprised a series of AI tools, the first of which included a quality-check convolutional neural network trained to identify that the image is of the retina and is of sufficient quality to be graded. Ungradable images were automatically identified and excluded from the analysis. The same convolutional neural network also classified whether the image belonged to the left or the right eye and whether the image was centred on the macula or optic nerve. Having passed through these steps, the image was passed onto an ensemble of grading a series of AI tools which were trained to grade retinopathy and maculopathy as separate entities.

The resultant grades were combined to produce a grading report 'per patient' (Figure 2). The grading AIs were designed to classify each image into one of three categories (Table 1): *non-sight-threatening* retinopathy; p*otentially referable disease*; and s*ight-threatening diabetic eye disease*.

In designing THEIA, an ensemble of convolutional neural networks was created and trained to find the optimal design. The best-performing architecture was then selected and went through another hyperparameter optimization process. THEIA was trained for 200 EPOCHS and the area under the receiver-operating curve was monitored.

Our basic premise in designing this primary triage AI system was that no individual with referable disease should be missed. When a human grader is reading a sequence of retinal images they issue a result based on the disease load visible across all the images acquired per eye; thus, on occasions, there will be only minimal disease visible in the individual macular and disc-centred images but the total load of disease across the two images represents referable disease. In contrast to the way the human grading team approaches a set of images, an AI system is only able to read and issue results one image at a time. Consequently, it is possible that referable disease would be missed by the AI system if the individual images of the sequence had minimal disease but the combined set had a total disease load that was greater than any of the individual images. Thus, to ensure that no disease was missed, an 'add-up' function was also selectively applied if the images recorded the same retinopathy scores. For example, if two images from the same eye were issued with an R2 grade, it would result in an R3 grade being issued for the eye. Issuing the maculopathy grade was more straightforward, being generated solely from the macula-centred image. Finally, and in recognition of the fact that, ultimately, it is the combined retinopathy and maculopathy grade across the two eyes that dictates the overall patient outcome, THEIA was designed to produce a per-patient outcome: non-sight-threatening, potentially referable, or sight-threatening (Figure 2).

In NZ, we have incorporated optical coherence tomography (OCT) into our community screening pathways; however, in the process of screening, small flecks of exudate can be easily overlooked, and consequently in those cases when it would most useful to have an OCT image, it is often not performed. To avoid this scenario, THEIA has been designed to alert the technician that sight-threatening maculopathy may be present and the technician is then invited to acquire an OCT whenever the AI detects that potentially sight-threatening maculopathy. This step was designed to ensure that an OCT is always available to the human grading team when the images are being manually graded.

2.6 | Assessing performance of the THEIA algorithm

As THEIA generates three possible outputs, namely, non-sight-threatening, potentially referable and sight-threatening disease, its performance is best described by way of 3×3 confusion matrices for both district health board datasets. To disclose the full performance of THEIA, these data are presented in Tables S1 to S9. However, the traditional metrics for describing the performance of screening methods are the sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) of the system to detect referable diabetic eye disease (combining both maculopathy and retinopathy together as a single entity) per patient. In order to create the bimodal outcome model needed to calculate these data, the outcomes of all images issued with a potentially referable (retinopathy and/or maculopathy) grade were combined with those issued with a sight-threatening (retinopathy and/or maculopathy) grade. This created two groups for the bimodal data analysis: non-referable disease (non-sight-threatening) and referable disease (potentially referable + sight-threatening). It should be noted that NPV and PPV depend on the prevalence of the diseased and healthy cases, but since the prevalence of referable disease (potentially referable + sight-threatening) was approximately 10% in our overall dataset, we do not anticipate this to be a statistical issue.

2.7 | THEIA clinical integration plan

THEIA's typical work cycle was <10 s per eye. Our proposed clinical implementation of THEIA into the NZ diabetic screening programme is shown in Figure 3.

3 | RESULTS

After excluding those images in which the either the maculopathy or retinopathy grade clinical reference standard was missing, the ADHB dataset included 160 585 images from 75 469 eyes, derived from 40 160 visits of unselected

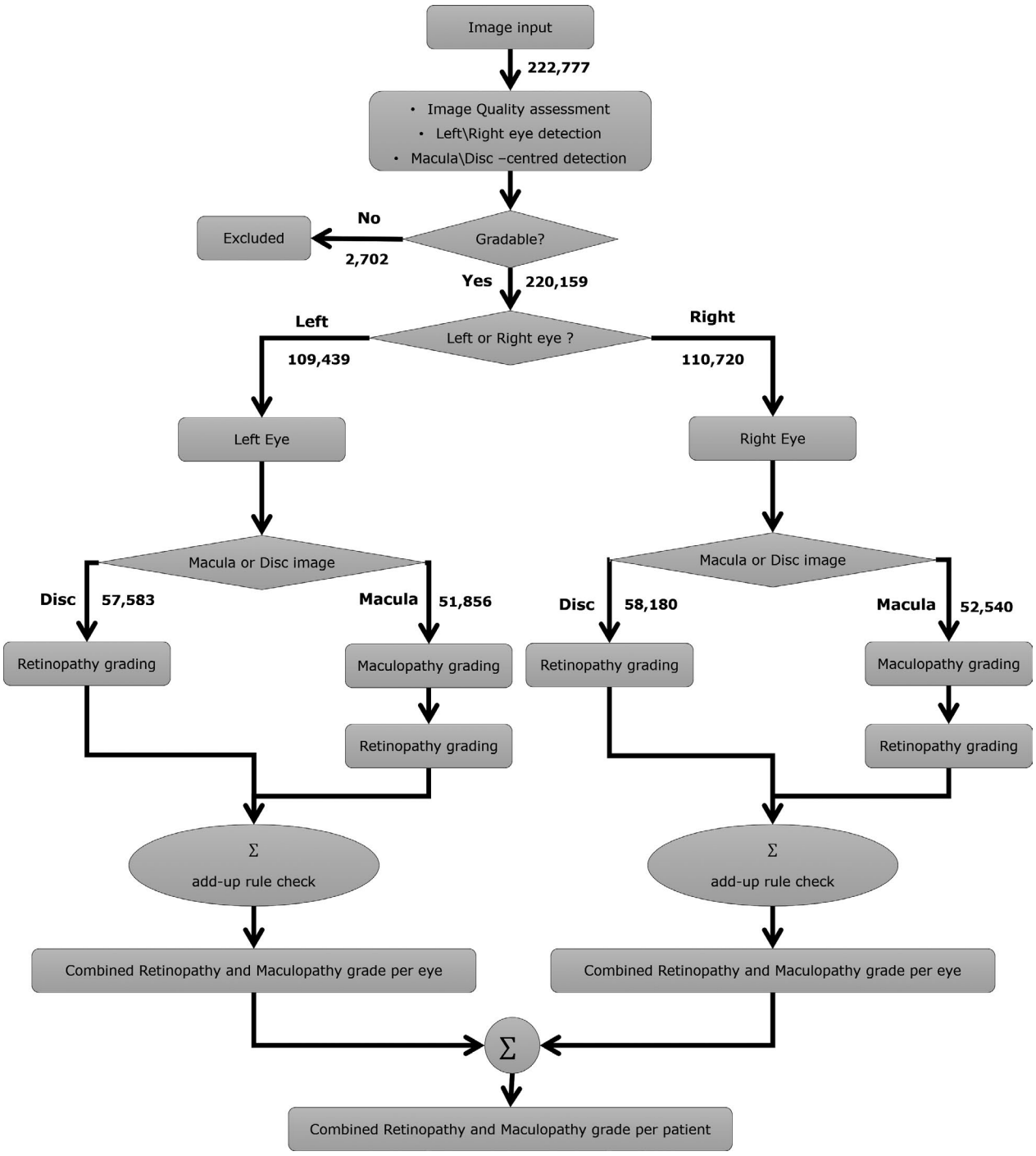

FIGURE 2 Flow chart of image processing pathway through THEIA

consecutive people with diabetes between 2009 and 2018, while the CMDHB dataset included 62 192 images from 37 147 eyes, derived from 23 683 visits of unselected consecutive people with diabetes between 2012 and 2018. In this project, only those eyes that had retinopathy *and* maculopathy grade results were included. Hence, there were 780 people with diabetes where the reference standard grading was missing within the ADHB dataset and 7858 people with diabetes where the reference standard grading was missing within the CMDHB dataset. The first component of the THEIA ensemble was designed to monitor the quality of the input image, and only when approved was an image allowed to process through the rest of the ensemble. During this process, images from just 2482 people with diabetes (4%) were rejected, the majority because of media opacities (cataract or corneal disease) or lid-generated artefacts (Fig. S1). This is very close to the current NZ clinical rates (image rejection rate based on CMDHB data: 2019/2020: 4.02%; 2018/2019: 5.66%).

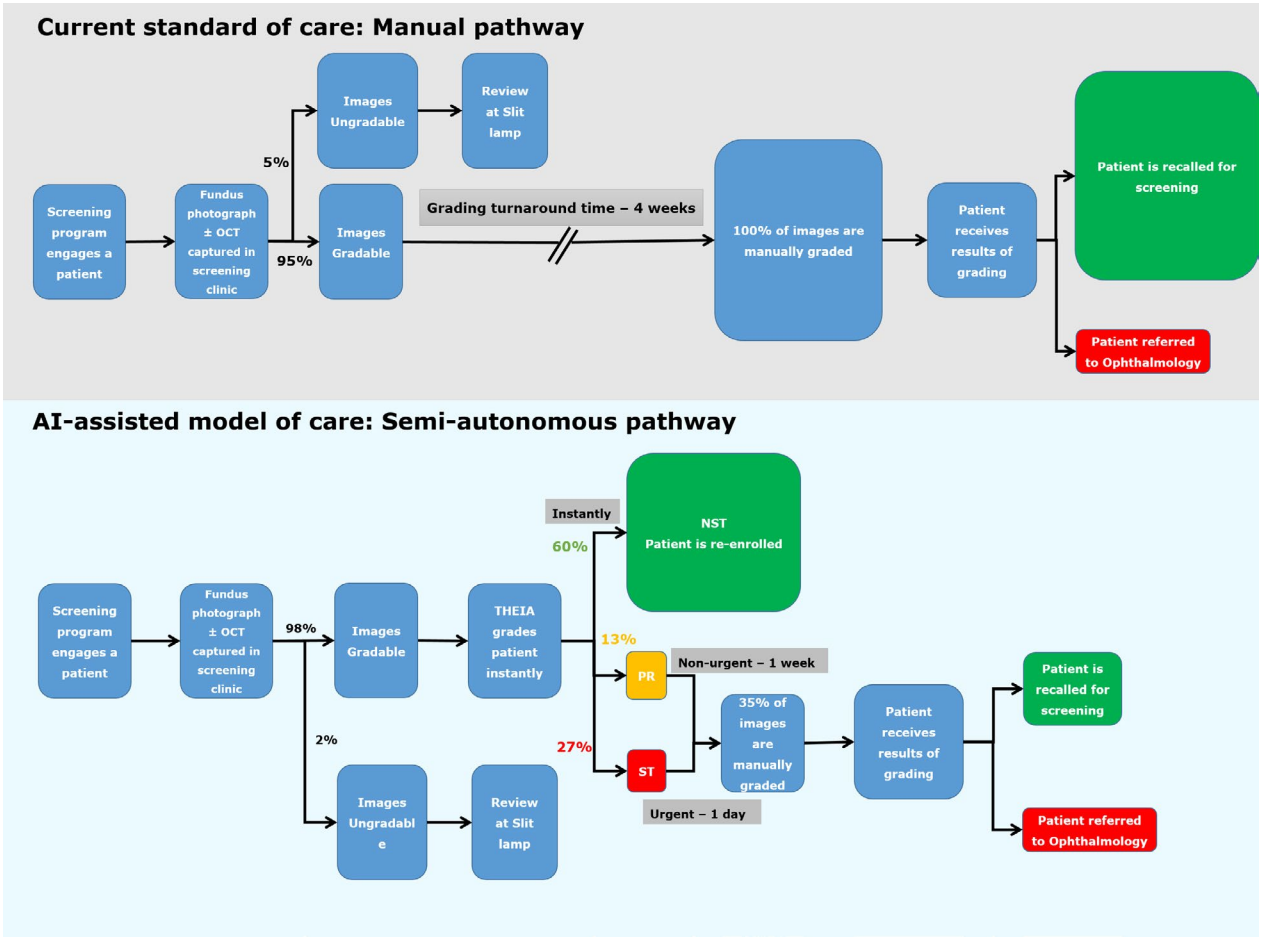

FIGURE 3 Proposed implementation of THEIA within the New Zealand diabetic screening programme, compared with the current model of care. AI, artificial intelligence; OCT, optical coherence tomography; PR, potentially referable; NST, non-sight-threatening; ST, sight-threatening

The baseline demographics of the two analysed district health board datasets are shown in Table 2. The retinopathy and maculopathy grades of the individuals whose photographs were included in this study are shown in Table 3. As is common with all DR screening programmes, over 50% (>30 000) of individuals in the CMDHB and ADHB datasets had no retinopathy. Fewer than 5% of individuals in the study population had sight-threatening DR. The two district health boards had slightly different distributions of disease severity, with a greater number of people with diabetes in the CMDHB population graded as having severe disease. A paired *t*-test on the ADHB and CMDHB dataset's disease distribution was performed and it was found that they were not significantly different ($P = 0.15$).

3.1 | Assessing the performance of THEIA per patient evaluated

In both validation datasets, the performance of THEIA to detect sight-threatening DR per patient was evaluated, based on retinopathy and maculopathy being treated as a single entity, compared to the reference standard (the result issued by the grading team at the time of screening). The calculated sensitivity, specificity and NPV for THEIA's performance to detect sight-threatening disease was 94%, 63% and 99.4%, respectively, in the ADHB validation dataset, and 95%, 61% and 99.4%, respectively, in the CMDHB validation dataset. Very similar results were observed if the data were analysed per eye rather than per patient. The full set of results are provided in Figs S2 and S3 and Tables S2 to S9.

3.2 | Grading load-saving estimate

A total of 65–70% of images screened with THEIA were issued with a non-sight-threatening grade at the time screening was performed. Arguably, the very high NPV means that images issued with a non-sight-threatening grade would no longer need review by the grading team. Hence, THEIA would reduce the image grading workload by approximately 65%.

TABLE 2 Demographics of the Auckland District Health Board and Counties Manukau District Health Board datasets

	ADHB	CMDHB	Combined
Number of images	160 585	62,192	222,777
Number of eyes	75 469	37 147	112 616
Number of visits	40 160	23 683	63 843
Number of unique cases	18 070	14 284	32 354
Mean (range) age, years	56 (7–98)	61 (7–104)	59 (7–104)
Women, %	46	48	47
Ethnicity, %			
NZ European	31	17	28
Māori/Cook Island Māori	12	20	16
Pacific	19	30	25
Other	36	28	28
Missing	2	5	3
Mean ± SD diabetes duration, years	9.1 ± 11.3	8.2 ± 7.3	8.8 ± 8.6

It should be noted that some images included missing labels (e.g. the reference standard for the maculopathy result) and were therefore not included in the analysis. Hence the number of eyes will be lower than expected (assuming that most people with diabetes will have two eyes in the screening programme).

3.3 | Audit of discordant grading

Using retrospective data for those designing large healthcare AI systems requires use of very large datasets. This approach ensures that, although discordant images will exist by random chance, this difference will be equally distributed across all outcomes.[16] In order to fully understand the performance of THEIA, we undertook an audit of those cases where the result generated by THEIA and the reference standard differed (Fig. S4). These errors fell into two groups: cases where THEIA genuinely produced the incorrect result and cases where THEIA generated the correct result and the original grading was incorrect. As expected by the inherent design of the system, the majority of instances where THEIA genuinely produced the incorrect result were false-positive errors. Of these, the most frequent in the maculopathy datasets were the foveal reflex and drusen being mislabelled as exudate and the most frequent false-positive errors in the retinopathy datasets were hypertensive flame haemorrhages being interpreted as blots, and small patches of resolving blot haemorrhage being interpreted as intraretinal microvascular abnormalities. The rate of false-negative errors was very low (0.2% and 0.5% in the two datasets); the commonest example being a small isolated patch of intraretinal microvascular abnormalities in the absence of any other DR being mislabelled by THEIA as non-sight-threatening. However, there were examples in both the maculopathy and retinopathy datasets where the original grading (reference standard) was wrong and THEIA was correct. The types of inconsistencies were very similar to those listed above.

4 | DISCUSSION

THEIA recorded a high sensitivity for detecting sight-threatening retinopathy of 95% and a very high NPV of 99.4% in the CMDHB and AHDB validation datasets, respectively. As THEIA categorized 65–70% of the images as non-sight-threatening disease, the time saved by removing this volume of cases from the human grading team's workload represents a significant cost saving for the programme. It would also ensure the consistency of grading nationally.

While direct comparison of different AI systems' performance is challenging as the data distribution and disease severity will be different between datasets used for training/validation, providing such comparison of the differing AI systems will give the reader a frame of reference by which to judge THEIA's performance. In comparison to recently published data, by design, THEIA exceeded EyeArt™ V2.0 in sensitivity (91%) and NPV (98%), but not in specificity (91%) and PPV (73%).[17] Similarly, THEIA exceeded the Pegasus™ system in sensitivity (84–93%) but not in specificity (85–94%).[18] The difference in the reported specificities and PPVs between THEIA, EyeArt™ and Pegasus™ is explained by the differing intended role for AI in DR screening. In contrast to other DR AI developers, we designed THEIA to maximize NPV so that it could safely perform primary triage, rapidly identifying the majority of people with diabetes with non-referable eye disease and thus reducing the burden on the human grading team. However, in order to minimize missed disease, we accepted a relatively low specificity, therefore, THEIA generates a number of false-positives. Whilst acknowledging that false-positives in any screening programme can cause significant anxiety to patients, THEIA was not designed to be a fully automated, stand-alone, DR screening platform. Instead, it has been designed to work within an established screening programme, allowing the NZ DR screening programme to transition to a semi-automated model, much as described by Xie *et al*.[19] recently. Within this model, THEIA is designed to perform the role of primary triage tool rapidly identifying those individuals with no disease who could be issued their results at the time of screening and those with sight-threatening disease whose images need urgent review by the human grading team. Whereas currently all patients have to wait a number of weeks for their screening result to be issued, implementing THEIA into the NZ screening programme will enable the majority of patients to be safely issued with a 'no significant disease present' result on the day of screening (Figure 3). The remainder will be passed to the grading team for a result to be issued. As the grading team no longer has to grade all the 'normal' images,

TABLE 3 Distribution of retinopathy and maculopathy grades (per eye) for Auckland District Health Board and Counties Manukau District Health Board datasets

ADHB					
Retinopathy, MoH grade	**Allocation, % (*n*)**	**THEIA grade, % (*n*)**	**Maculopathy, MoH grade**	**Allocations, % (*n*)**	**THEIA grade, % (*n*)**
R0	53 (84 427)	Non-sight-threatening	M0	68 (10 8633)	Non-sight-threatening
R1	33 (53 213)	97.1 (155 825)	M1	13 (20 204)	96.5 (154865)
R2	11 (18 185)		M2	16 (25 999)	
R3	2 (2955)	Potentially referable	M3	2 (2374)	Potentially referable
		1.8 (2995)			1.5 (2374)
R4	1 (1515)	Sight-threatening	M4	2 (3064)	Sight-threatening
R5	0.2 (250)	1.1 (1765)	M5	0.1 (141)	2.0 (3205)
CMDHB					
Retinopathy, MoH grade	**Allocations, % (*n*)**	**THEIA grade, % (*n*)**	**Maculopathy, MoH grade**	**Allocations, % (*n*)**	**THEIA grade, % (*n*)**
R0	55 (34 088)	Non-sight-threatening	M0	48 (29 996)	Non-sight-threatening
R1	21 (13 318)	96.3 (58 837)	M1	14 (8586)	95.1 (119 017)
R2	18 (11 431)		M2	32 (19 485)	
R3	4 (2181)	Potentially referable	M3	2 (1437)	Potentially referable
		2.4 (3031)			1.7 (2113)
R4	2 (979)	Sight-threatening	M4	4 (2564)	Sight-threatening
R5	0.3 (164)	1.3 (1570)	M5	0.1 (90)	3.1. (3870)

Abbreviations: ADHB, Auckland District Health Board; CMDHB, Counties Manukau District Health Board. Numbers in brackets are indicative of the number of images in each category: R0, no retinopathy; R1, minimal non-proliferative retinopathy; R2, mild non-proliferative retinopathy; R3, moderate non-proliferative retinopathy; R4, severe non-proliferative retinopathy; R5, proliferative retinopathy; M0, no maculopathy; M1, at least one microaneurysm beyond 1 disc diameter of the fovea but within the macular arcades; M2 at least one microaneurysm within 1 disc diameter of the fovea; M3, exudate beyond 1 disc diameter of the fovea, but within the macular arcades; M4, exudate within 1 disc diameter of the fovea; M5, similar to M4, but with vision loss.[14]

individuals will have their results issued more rapidly than was previously possible. Moreover, the design of THEIA means that all individuals who may have sight-threatening referable disease are automatically flagged, allowing these images to be reviewed urgently by the human grading team. Ultimately, only a small number of patients will need ongoing referral to an ophthalmology clinic for review and no patient will be inconvenienced by being asked to attend for ophthalmology review on the basis of THEIA generating a false-positive result.

Internationally, there are very few trained neural networks that have (1) used publicly available datasets for training, (2) externally validated their AI, and (3) issued a grade for both retinopathy and maculopathy.[16,20–23] Verbraak *et al*.[23] recently published results on the performance of the IDX–DR-EU 2.1, where they assessed multiple images grading the eye disease into retinopathy and maculopathy, and then collapsed the grades down into a binary 'vision-threatening' versus 'non-vision-threatening' matrix. The reported performance of this device is very similar to THEIA.

A confounding factor which makes direct comparisons difficult between AI systems is that many use highly curated data and this ultimately risks reducing both the utility and impact of the AI system.[3,5,6,20,23–27] IDX-DR-EU2.1 was unable to issue a grade in approximately 20% of cases, compared to THEIA which was able to analyse 96% of all people with diabetes in our uncurated datasets. Crucially, THEIA also includes a broad age range of people with diabetes (aged 7–104 years) of various ethnicities. Where demographic characteristics have been reported, many developers have excluded individuals who were aged <40 years[20] and this risks the AI being unable to deal with the highly reflective internal limiting membrane within the macular, a feature which is very prominent in retinal images obtained from young individuals.

To date, most published DR screening AI systems have only graded a single macular-centred image.[6,20,22,24–28] However, it is recognized that a single image cannot be relied on for accurate grading and an AI system that is trained to read only a single image per eye risks under-grading the peripheral retinopathy. In contrast, THEIA reads all images from any given patient. It then collates the combined

maculopathy and retinopathy grades to confer the ultimate patient outcome. As previously outlined, THEIA also uses a novel post-image analysis processing 'add-up function' to ensure that no disease is missed. Whilst this is approach will reduce the likelihood of referable disease being missed, it will reduce the specificity of the system by generating a number of false-positives. However, as stated above, we are confident that a high false-positive rate in an AI system that is designed to perform a specific role within an established screening programme will not ultimately disadvantage the patient. Finally, to address the 'black box' issue of other AI systems, THEIA generates 'attention maps' for retinopathy and maculopathy which can be viewed by a clinician post-diagnosis (Fig. S3).

While there has been significant interest in developing DR grading AI systems,[29] very few have been trained to specifically grade diabetic maculopathy as a separate entity;[6,16,20,22,24,25,27] this is despite diabetic maculopathy being the commonest reason for ophthalmology referral.[30] One of the reasons for this anomaly lies in the intrinsically challenging nature of detecting and grading maculopathy with surrogate markers of tissue oedema, macular exudates and perifoveal micro-aneurysms, being used to identify disease. To improve the detection of suspected maculopathy, if THEIA registers that sight-threatening maculopathy may be present, it prompts the technician taking the photographs to perform an OCT scan before completing the imaging sequence.

A dilemma that is inherent in all healthcare AI systems that use large prospective datasets for training and internal validation is the possibility that the original results issued at the timing of grading are incorrect. Such errors will introduce systemic bias into the system which is one of the reasons why AI systems trained on small datasets often fail to perform when tested on external datasets.[31] Whilst the presence of a similar bias in THEIA will only be revealed by a prospective double-blind clinical trial, we believe that the results of the present study remain valid for three reasons. Firstly, the reference standards used for the original set of training images were carefully re-graded by the lead ophthalmologist. Secondly, to mitigate against the possibility that one of the district health board's grading processes had a consistent bias, we used two very large datasets derived from district health boards that have different grading teams. Thirdly, during the development of THEIA, all discordant results between the AI and the reference standard were constantly audited and the AI retrained accordingly. Nevertheless, despite all these safeguards, human error will still occur and thus one would still expect there to be some discordance between the final results generated by the AI system and the reference standard. The results of our internal audit suggest that the rate of these errors was low at <0.5%.

Whilst many of the AI systems that have been developed to date have been designed to be fully automated, stand-alone screening systems, we have taken a different approach and have instead designed a bespoke AI system that will allow the NZ diabetic eye screening programme to transition to a semi-automated system, with the AI system playing a specific role as a rapid primary triage tool. Xie *et al*.[19] have recently published a model-based cost-minimization analysis of three DR screening models in Singapore, comparing the current human-centric screening model to both a semi-automated model 'backed up' by secondary human confirmatory diagnoses, and a fully automated model without any human assessment. Their findings appear to vindicate our approach as the authors concluded that the model utilising a semi-automated AI system was the most cost-effective, costing ~6% less than the fully automated model, and 20% less than the manual model.

The present study has several limitations. Perhaps the most important clinical limitation of THEIA currently is that it cannot identify other common eye diseases, such as glaucomatous optic neuropathy and age-related macular degeneration. However, from the inception of our screening programme, we have routinely recorded all non-diabetic eye disease detected during screening. A recent analysis of these data by the authors (manuscript submitted) revealed that, of these, only hypertensive retinopathy and macular degeneration are sufficiently important to justify detection during routine diabetic eye screening. Although, in this current iteration of THEIA, these diseases cannot be identified by name, the 'miss no disease' design of THEIA means that most of the signs of hypertensive retinopathy and most cases of macular degeneration will be picked up and flagged up as potentially referable or sight-threatening. From the perspective of an AI system, the low number of cases with referable DR is problematic as this creates an imbalanced dataset which can affect the quality of trained AI. The retrospective nature of the study is also an issue, but this approach is necessary in unselected, unbiased, real-world datasets. Finally, although this first iteration of THEIA has been developed from the two largest and demographically diverse district health boards in NZ, its generalizability remains to be tested in a prospective study.

In conclusion, THEIA provides high sensitivity for detecting sight-threatening retinopathy and a corresponding NPV of 99.4% in primary grading of retinal images within the context of the NZ Ministry of Health guidelines for DR screening. This first iteration of THEIA accurately identifies people with diabetes with non-referable disease, removing such people with diabetes from the manual grading workload, a group which represents over 65% of the images. As such, it will probably generate significant efficiencies for the national screening programme, and will deliver greater consistency among regions. Although this first iteration of THEIA has been developed for the NZ context, with appropriate tailoring, this system has potential for application in

other healthcare systems which have, or intend to develop, a structured DR screening programme.

COMPETING INTERESTS

E.V. and D.S. are co-founders of Toku Eyes®, which is a start-up from the University of Auckland, looking into commercialization of this artificial intelligence system (THEIA™) in NZ. The remaining authors have no conflicts of interest to declare.

ACKNOWLEDGEMENTS

We wish to acknowledge Stephanie Emma, Coordinator of the Diabetes eye screening programme at the CMDHB for providing the data from South Auckland.

ORCID

E. Vaghefi https://orcid.org/0000-0002-9482-3168

REFERENCES

1. Nørgaard MF, Grauslund J. Automated screening for diabetic retinopathy–a systematic review. *Ophthalmic Res*. 2018;60:9-17.
2. Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. 2016;44:260-277.
3. Bellemo V, Lim G, Rim TH, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diabetes Rep*. 2019;19:72.
4. Lee A, Taylor P, Kalpathy-Cramer J, Tufail A. Machine learning has arrived!. *Ophthalmology*. 2017;124:1726-1728.
5. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
6. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1:e35-e44.
7. In: Weir J, ed. *New Zealand's population reflects growing diversity*. 2019. News published by Stat New Zealand on domestic population growth. Available at https://www.stats.govt.nz/news/new-zealands-population-reflects-growing-diversity Last accessed 20 September 2020.
8. Ballantyne A. Adjusting the focus: a public health ethics approach to data research. *Bioethics*. 2019;33:357-366.
9. Ballantyne A, Style R. Health data research in New Zealand: updating the ethical governance framework. *NZ Med J*. 2017;130:64-71.
10. Raine SC, Kukutai T, Walter M, Figueroa-Rodrigues OL, Walker J, Axelsson P. Indigenous data sovereignty. The State of Open Data.:300. 2019. Available at https://library.oapen.org/bitstream/handle/20.500.12657/24884/The_State_of_Open_Data_9781928331957_web.pdf?sequence=1#page=315
11. Silveira T, Hudson M. GOVERNANCE OF MAORI DATA. 2018. Available at http://www.maramatanga.ac.nz/sites/default/files/project-reports/Silveira%2C%20Tumanako%20-%2017INT22%20-2017%20-%20PDF%20Report.pdf
12. Xie L, Yang S, Squirrell D, Vaghefi E. Towards implementation of AI in New Zealand national screening program: Cloud-based, Robust, and Bespoke. *bioRxiv* 2019:823260.
13. Scott C. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*. 2012;6:958-992.
14. Health NZMo. Diabetic Retinal Screening, Grading, Monitoring and Referral Guidance. In: Health NZMo, ed. *New Zealand Ministry of Health - publications*. 2 edn. : www.health.govt.nzwww.health.govt.nz 2016.
15. Graham B. Kaggle diabetic retinopathy detection competition report. *University of Warwick* 2015.
16. Ting DSW, Cheung CY-L, Lim G, 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;318:2211-2223.
17. Bhaskaranand M, Ramachandra C, Bhat S, et al. The Value of Automated Diabetic Retinopathy Screening with the EyeArt System: A Study of More Than 100,000 Consecutive Encounters from People with Diabetes. *Diabetes technology & therapeutics*. 2019.
18. Rogers T, Gonzalez-Bueno J, Franco RG, et al. Evaluation of an AI System for the Detection of Diabetic Retinopathy from Images Captured with a Handheld Portable Fundus Camera: the MAILOR AI study. *arXiv preprint arXiv:190806399* 2019.
19. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *The Lancet Digital Health* 2020.
20. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*. 2016;316:2402-2410.
21. Li Z, Keel S, Liu C, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes Care*. 2018;41:2509-2516.
22. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57:5200-5206.
23. Verbraak FD, Abramoff MD, Bausch GC, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care*. 2019;42:651-656.
24. Gulshan V, Rajan RP, Widner K, et al. Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA Ophthalmol*. 2019;137:987-993.
25. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264-1272.
26. Ramachandran N, Hong SC, Sime MJ, Wilson GA. Diabetic retinopathy screening using deep neural network. *Clin Exp Ophthalmol*. 2018;46:412-416.
27. Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep Learning vs. Human Graders for Classifying Severity Levels of Diabetic Retinopathy in a Real-World Nationwide Screening Program. *arXiv preprint arXiv:181008290* 2018. Available at https://arxiv.org/abs/1810.08290.
28. Voets M, Møllersen K, Bongo LA. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *arXiv preprint arXiv:180304337* 2018. Available at https://arxiv.org/abs/1803.04337.
29. Ou WC, Wykoff CC. The Promise of Deep Learning in Retina. 2018. https://www.retina-specialist.com/article/the-promise-of-deep-learning--in-retina-1-1

30. Ruta L, Magliano D, Lemesurier R, Taylor H, Zimmet P, Shaw J. Prevalence of diabetic retinopathy in Type 2 diabetes in developing and developed countries. *Diabet Med*. 2013;30:387-398.
31. Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS One*. 2019;14:e0217541.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Vaghefi E, Yang S, Xie L, et al. THEIA™ development, and testing of artificial intelligence-based primary triage of diabetic retinopathy screening images in New Zealand. *Diabetic Medicine*. 2021;38:e14386. https://doi.org/10.1111/dme.14386